Applied Computing and Intelligence
2(1): 1–31
DOI: 10.3934/aci.2022001
Received: 11 March 2022
Revised: 29 March 2022
Accepted: 30 March 2022
Published: 31 March 2022

https://www.aimspress.com/journal/aci

Survey

A comprehensive survey of zero-shot image classification: methods, implementation, and fair evaluation

Guanyu Yang[1], **Zihan Ye**[1], **Rui Zhang**[2] **and Kaizhu Huang**[3,*]

[1] Department of Intelligent Science, Xi'an Jiaotong-Liverpool University, SIP, Suzhou, 215123, China

[2] Department of Foundational Mathematics, Xi'an Jiaotong-Liverpool University, SIP, Suzhou, 215123, China

[3] Institute of Applied Physical Sciences and Engineering, Duke Kunshan University, Kunshan, 215316, China

* **Correspondence:** kaizhu.huang@dukekunshan.edu.cn

Academic Editor: Chih-Cheng Hung

Abstract: Deep learning methods may decline in their performance when the number of labelled training samples is limited, in a scenario known as few-shot learning. The methods may even degrade the accuracy in classifying instances of classes that have not been seen previously, called zero-shot learning. While the classification results achieved by the zero-shot learning methods are steadily improved, different problem settings, and diverse experimental setups have emerged. It becomes difficult to measure fairly the effectiveness of each proposed method, thus hindering further research in the field. In this article, a comprehensive survey is given on the methodology, implementation, and fair evaluations for practical and applied computing facets on the recent progress of zero-shot learning.

Keywords: zero-shot; classification; inductive; transductive; survey; implementation

1. Introduction

In the field of computer vision, deep learning methods have made great achievements in both applied computing and machine intelligence. Remarkably, deep learning attains unprecedented success in image classification. Exploiting many powerful deep neural networks, machines can perform at a level close to or even beyond that of humans in many applications as long as sufficient labelled samples are provided [29, 78, 89]. However, the conventional deep neural network models rely on many important factors in order to achieve excellent performance. Typically, deep neural networks require a huge number of labelled samples for training, whilst massive sample collection and labelling may unfortunately be difficult, time-consuming, or even impossible in many cases.

In fact, not in line with deep neural networks' high demanding on data, there are many scenarios which are commonly seen in practice:

- **Large target size**. Human beings could distinguish around 3,000 basic-level classes [6], and each basic class could be expanded as subordinate ones, such as dogs in different breed [115]. Such a huge number of categories makes it infeasible to construct a task where each category has a sufficient number of labelled samples.
- **Rare target classes**. Some tasks suffer from rare classes for which the corresponding samples are difficult to be obtained, such as fine-grained classification over flowers and birds [13, 46] or medical images corresponding to certain specific situation [11].
- **Growing target size**. The target set for some tasks changes rapidly, with candidate classes increasing over time, such as detection of new events in newly collected media data [10], recognizing the brand of a product [61] or learning some writing styles [35].

In those scenarios, re-training a deep neural network model over target classes appears not very feasible. Fine-tuning the trained model might be tractable only if some of the labelled target samples could be obtained. To overcome such restrictions, zero-shot learning, earlier called zero-data learning, is set up to simulate the learning capacity of human beings [45]. Assuming a child is equipped with knowledge including the shape of the horse, the concept of stripes, and colours of black and white, once being told that zebra looks like a horse covered in black and white stripes, the child has a good chance of recognizing a zebra even if seeing it for the first time [19]. Figure 1 demonstrates a schematic graph for the efficient learning process that situations are also similar in zero-shot learning. Based on the auxiliary information used to describe each category and some corresponding samples, a model can be trained to construct the correlation between samples and the auxiliary information, thus enabling to extend classification on unseen categories, based on their correlation as well as the auxiliary information.

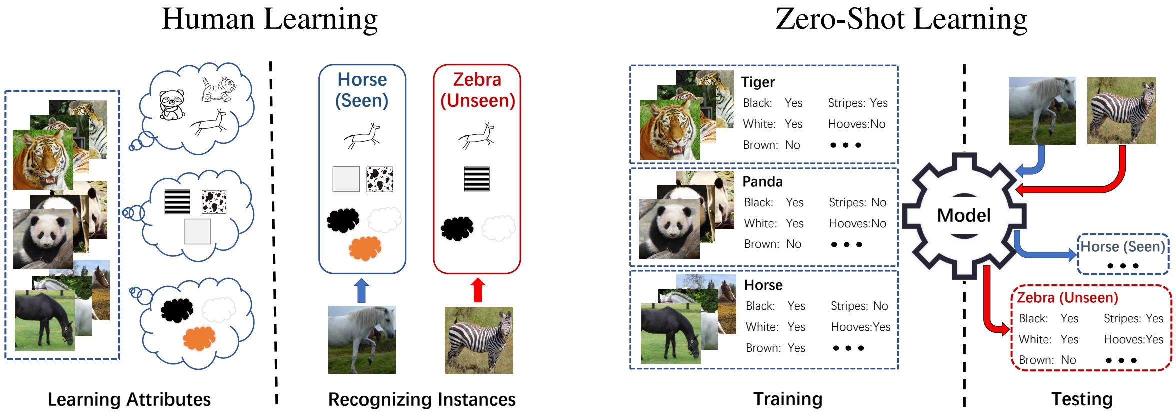

Figure 1. Examples for learning processes.

In this article, we present an overview of image classification in zero-shot learning including its relevant definitions, learning scenarios, and various methodologies. While we properly structure each

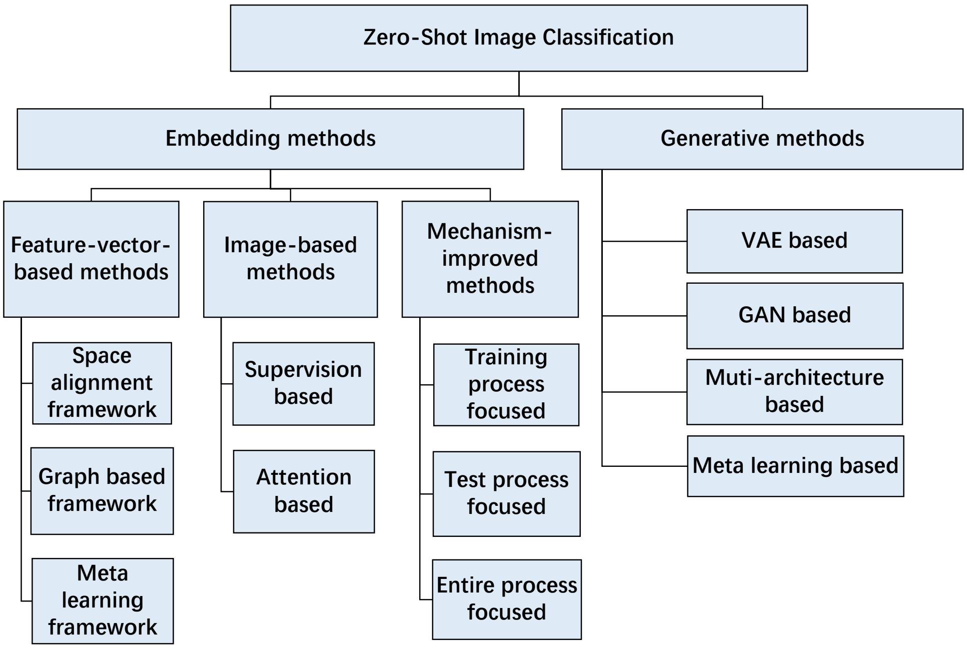

Figure 2. The taxonomy structural diagram for Zero-Shot image classification methods.

part and summarize each family of methods with illustration, visualization, and tables, we put one main focus of this work on sorting out the implementation details, such as commonly used benchmarks, and diverse experiment settings so as to offer a more practical guidance to researchers in the area. In the end, the comparison results of various representative methods are collected on a number of benchmarks, aiming to provide a fair and objective reference for evaluating different methods.

Compared to the recently-presented surveys [73, 99], our paper shows three major differences. First, our work introduces most recently published important methods, as more seminar works and even breakthroughs emerged recently, thus reflecting a more timely and comprehensive review. Second, based on model components, training strategies, and learning objectives, we provide a more detailed hierarchical classification for zero-shot image classification methods. Third, we put one main focus of our survey on comparing different methods from the perspective of implementations, thus offering practical guidelines for applying zero-shot learning on real scenarios.

2. Overview of zero-shot learning

To describe the zero-shot classification task precisely, we will first review and explain some commonly used terms and notations in this section, then focus on introducing the zero-shot image classification methods which employ semantic descriptions as auxiliary information in the next two sections. Based on the design of the information extractor, we classify the current methods into two main categories: *embedding methods* and *generative methods*, and propose a taxonomy structure for these methods as shown in Figure 2. For simplicity of expression, all subsequent references to zero-shot learning refer to the image classification task under this domain.

2.1. Auxiliary information

In zero-shot learning, the target classes without corresponding training samples are named as unseen classes, whilst the classes with labelled samples during training are called seen classes. Due to the absence of training samples for unseen classes, auxiliary information is essential for constructing the cognitive concepts of unseen categories. The space of such auxiliary information should contain enough information to distinguish all classes. In other words, for each class, corresponding auxiliary information should be unique and sufficiently representative to guarantee that an effective correlation between the auxiliary information and the samples can be learned for classification. Since zero-shot learning is inspired from the human efficient learning process, semantic information has become the commonly dominant auxiliary information [46, 44, 95]. Similar to the feature space for image processing, there is also a corresponding semantic space holding numeric values in zero-shot learning. To obtain such semantic space, two different kinds of semantic sources, attributes and textual descriptions, are mainly leveraged.

Attribute. Attribute is the earliest and most commonly used source of semantic space in zero-shot learning [43, 45, 99]. As a kind of human-annotated information, attribute contains precise classification knowledge though its collection might be time-consuming. Considering an attribute as a word or phrase introducing a property, one can build up a list of attributes. By combing these attributes, all the seen and unseen classes can be described. Moreover, these combined descriptions should be different for each class. Then the vectors, holding binary values 0 and 1 with sizes equal to the number of the attributes, form a semantic space where each value denotes whether the described class is equipped with the corresponding attributes or not. In other words, the attribute vectors for all the classes share the same size, and each dimension of the vector denotes a specific property in a settled order. For example, in animals recognition, one attribute could be *stripe*. Value 1 in the dimension of *stripe* of the attribute vector means that the described animal is with stripes [43]. Suppose there are only 3 attributes: *black*, *white*, and *stripes*, then the attribute vectors describing classes *panda*, *polar bear* and *zebra* should be something like [1, 1, 0], [0, 1, 0] and [1, 1, 1] respectively. However, since an attribute vector is designed to describe the entire class, it might be imprecise to use binary values only. The diversity of individuals within each class may lead to a mismatch between the sample and attributes. Taking again the animal recognition as an example, we can see horses might be also in pure black and pure white. If the attribute values of both *black* and *white* equal 1 for the class *horse*, then the black horse samples are contradictory to the attribute *white*, so are the white horses to the *black*. Therefore, instead of taking the binary value, it makes more sense to employ continuous values indicating the degree or confidence level for an attribute. It is shown in [2] that adopting the average value of the voting results or the proportion of the samples corresponding to an attribute leads to better classification performance. Additionally, the relative attribute measuring the degree of attribute among classes is also suggested [71].

Text. Instead of using human-annotated attributes, descriptions of a class such as the name or definition could also be considered as the source to construct a semantic space. However, it is not straightforward to transform the unstructured textual information into representative real values. When the class name is exploited as the semantic source without any external knowledge, the contained information might be far from enough for achieving good classification among images. In this case, pre-trained word embedding models borrowed from natural language processing could embed the class names to some representative word vectors and form a meaningful semantic space. Specifically, the semantic similarity of two vocabularies can be approximately measured by the distance between the two corresponding embedded vectors, thus the similarity knowledge contained in the training text corpora (for constructing the word embedding models) could be adopted for classification. In the existing methods, Word2Vec [3, 69, 96, 103] and GloVe [3, 103, 58] pre-trained on English language Wikipedia [85] are two commonly used embedding models for class name sources. Such semantic similarity measure space can also be constructed via the knowledge in terms of ontology. An example is to adopt the hierarchical embedding from a large-scale hierarchical database WordNet [3]. The keyword is another optional semantic source. The descriptions of classes are collected through databases or search engines to extract keywords. Consequently, the binary occurrence indicator [74] or frequencies [3] in Bag-of-Words, or transformed term frequency–inverse document frequency features [13, 46, 14] can construct such semantic vectors. The description in the form of paragraph could also be used as a semantic source. For example, visual descriptions in the form of ten single sentences are collected for images in [76]. After that, the text encoder model is utilized to return the required semantic vectors. This kind of semantic source contains more information as well as more noises.

Other auxiliary information. In addition to the semantic source, other types of supporting information also exist. That kind of information is often employed simultaneously with semantic information to assist the model in extracting more effective classification knowledge. For instance, hierarchical labels in taxonomy are introduced to provide additional supervision of classification [79, 107]; the human defined correlation between attributes [32] capturing the gaze point of each sample is adopted as the attention supervision to improve the attention module producing more representative feature maps [58]; Some of these information may not provide sufficient knowledge to accomplish the entire classification task. However, they can be regarded as the supplementary of semantic information which may better construct cognitive concepts of unknown categories.

2.2. Learning scenarios

In conventional image classification tasks, due to the differences in the distribution of instances between the training and test sets, the trained model does not perform as well during the test as it does on the training set. This phenomenon is also present in zero-shot learning, and even more severe owing to the disjoint property of seen and unseen classes. Such differences in the distribution between seen and unseen classes are called domain shift [18]. Moreover, the poor model performance is termed as class-level over-fitting [120].

To address this challenge, by effectively employing classification knowledge from samples and auxiliary information, researchers have proposed various methods of introducing knowledge at different stages (including training and testing). As a result, the implementation scenarios become diverse. Both sample space and auxiliary information space can be defined in zero-shot learning, according to which we can divide the scenarios accordingly. In general, from the perspective of the training stage, the task can be divided into three scenarios, namely inductive, semantic transductive, and transductive, which are defined as follows:

- **Inductive zero-shot learning**. Only labelled training samples and auxiliary information of seen classes are available during training.
- **Semantic transductive zero-shot learning**. Labelled training samples and auxiliary information of all classes are available during training.
- **Transductive zero-shot learning**. Labelled training samples, unlabelled test samples, and auxiliary information of all classes are available during training.

From the definition, the inductive zero-shot learning represents the most severe learning scenario because both the target classes and instances are unknown. Models trained in this scenario are more likely to suffer from class-level over-fitting. In comparison, models trained in the rest two transductive scenarios share a clear learning objective since the classification knowledge is guided by the unseen information. However, these trained models will not generalize to new unseen classes as well as the models trained in the inductive scenario [99].

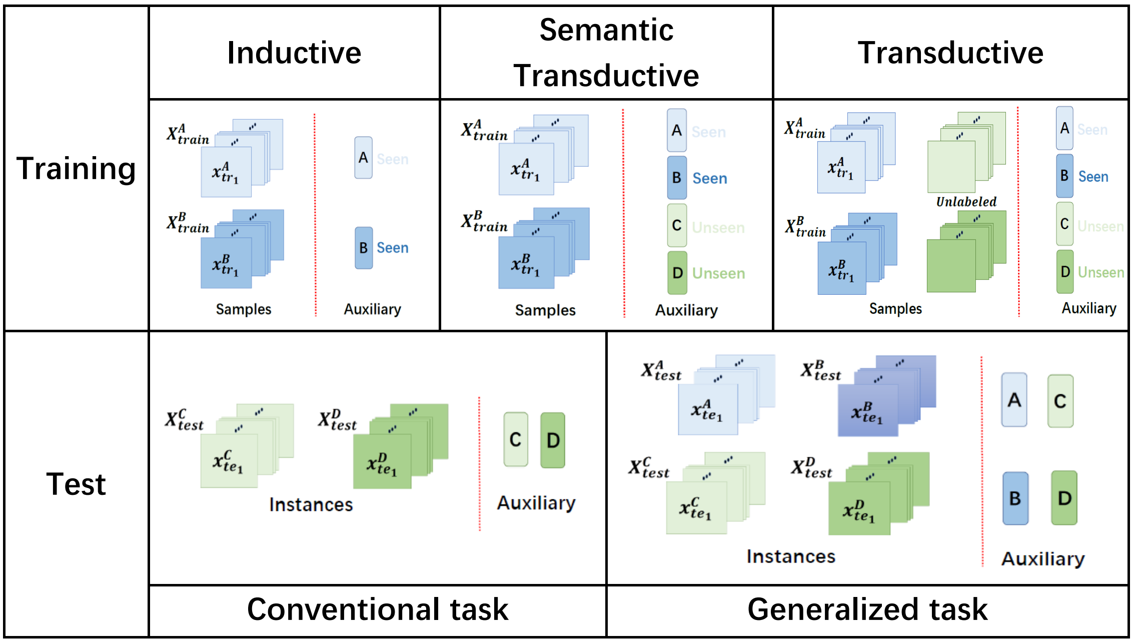

Figure 3. Schematic diagrams of utilizing data for different scenarios in training and test.

When the zero-shot problem was first proposed in the early stage, researchers focused only on achieving good classification on unseen classes, which is known as conventional Zero-Shot Learning. Later, it was found that the classification of the unseen classes would suffer from a devastating blow once the seen categories were also included as candidates for classification. In other words, the early proposed models could not distinguish well between seen and unseen categories and thus fail to construct the cognition concepts of new classes. Consequently, a more challenging task called Generalized Zero-Shot Learning attracts much attention, which requires classifying both seen and unseen classes [9]. The original intention of zero-shot learning is to simulate the human process of constructing the cognition concept of classes from learned knowledge and supporting information in the absence of samples. Since the constructed cognitive concepts can be evaluated accurately only if the unseen and seen classes can also be correctly distinguished, the focus of current works has shifted to the generalized one. Figure 3 shows the schematic of different scenarios in training and test, where the combination of the different scenarios forms six common settings.

2.3. Problem definitions

In zero-shot learning, each sample is originally designed as an image containing certain specific objects in a tensor form holding value for each pixel. To ensure more convenient implementation, the visual features extracted by a pre-trained deep neural network are commonly regarded as the samples instead of using the image. For a rigorous presentation, here we take the entire image as the input sample in our article. Assuming there are totally N samples from K classes, we denote $\boldsymbol{X} = \boldsymbol{X}^S \cup \boldsymbol{X}^U$ as the set of all the image samples from both seen and unseen classes, and $\mathbf{F}(\cdot)$ as a feature extractor for obtaining the feature $\mathbf{F}(x_i)$ of the image x_i. Similarly the corresponding label set could be denoted as $\boldsymbol{Y} = \boldsymbol{Y}^S \cup \boldsymbol{Y}^U$, and $y_i = k$ indicates that sample x_i belongs to the kth-class. The set of the auxiliary information is denoted as $\boldsymbol{A} = \boldsymbol{A}^S \cup \boldsymbol{A}^U$ which contains K vectors where each vector a_k stands for the auxiliary information of the kth-class. Here let K^S and K^U indicate the number of seen and unseen classes respectively and the first K^S classes represented in $\boldsymbol{A}$ are assumed as the seen ones for convenience. Note that the seen and unseen classes are disjoint, which means $\boldsymbol{X}^S \cap \boldsymbol{X}^U = \boldsymbol{Y}^S \cap \boldsymbol{Y}^U = \boldsymbol{A}^S \cap \boldsymbol{A}^U = \emptyset$. As partial of seen class samples are adopted as test instances which should not participate in the training process, the seen sets of the samples and labels are further consistently divided into training and test sets as $\boldsymbol{X}^S = \boldsymbol{X}^S_{tr} \cup \boldsymbol{X}^S_{te}$ and $\boldsymbol{Y}^S = \boldsymbol{Y}^S_{tr} \cup \boldsymbol{Y}^S_{te}$. Specifically, both of the train and test seen sets should cover all the K^S seen classes. Since there are three scenarios for the training process, the training set $\boldsymbol{D}_{tr} = \{\boldsymbol{X}_{tr}, \boldsymbol{Y}_{tr}, \boldsymbol{A}_{tr}\}$ can be respectively defined for the inductive, semantic transductive, and transductive scenarios in the three forms as $\boldsymbol{D}^I_{tr} = \{\boldsymbol{X}^S_{tr}, \boldsymbol{Y}^S_{tr}, \boldsymbol{A}^S\}$, $\boldsymbol{D}^{ST}_{tr} = \{\boldsymbol{X}^S_{tr}, \boldsymbol{Y}^S_{tr}, \boldsymbol{A}\}$ and $\boldsymbol{D}^T_{tr} = \{\boldsymbol{X}^S_{tr} \cup \boldsymbol{X}^U, \boldsymbol{Y}^S_{tr}, \boldsymbol{A}\}$. For the test set $\boldsymbol{D}_{te} = \{\boldsymbol{X}_{te}, \boldsymbol{Y}_{te}, \boldsymbol{A}_{te}\}$, it can also be defined in two forms as $\boldsymbol{D}^C_{te} = \{\boldsymbol{X}^U, \boldsymbol{Y}^U, \boldsymbol{A}^U\}$ for conventional task and $\boldsymbol{D}^G_{te} = \{\boldsymbol{X}^U \cup \mathbf{X}^S_{te}, \boldsymbol{Y}^U \cup \boldsymbol{Y}^S_{te}, \boldsymbol{A}\}$ for generalized task respectively. With these definitions, the target of zero-shot learning can be represented to train an information extractor $\mathbf{M}$ (containing the feature extractor $\mathbf{F}(\cdot)$) with a settled or a learnable classifier $\mathbf{C}$ on the training set $\boldsymbol{D}_{tr}$ to achieve classification on $\boldsymbol{X}_{te}$.

3. Embedding methods

In the embedding methods, the information extractor $\mathbf{M} = \{\theta(\cdot), \ \phi(\cdot)\}$ is designed as a union of embedding functions $\theta(\cdot)$ and $\phi(\cdot)$. The aim of these extractors is to find the proper embedding spaces for both visual samples and auxiliary information so that the trainable or settled classifier $\mathbf{C}$ can achieve class recognition among the target space. From the perspective of the learning objective, we further classify the existing embedding methods as: (1) *feature-vector-based*, (2) *image-based*, and (3) *mechanism-improved* methods.

3.1. Feature-vector-based methods

Considering the limitation of the sample size and the latent distribution differences between the samples of the unseen and seen classes, the most easily associated and appropriate visual feature space is the learned space in large-scale conventional image classification tasks. Fair data splits and extracted features for several benchmarks are discussed and evaluated in [104]. The feature vector space learned by the deep residual network called ResNet101 [26] over a benchmark dataset ImageNet [12] is commonly selected in the implementations. Based on the fixed feature extractor $\mathbf{F} = \mathbf{F}_f$, feature vectors $\mathbf{F}_f(\boldsymbol{X})$ are regarded as the visual samples and the insight of the feature-vector-based methods is to design embedding functions or classifiers trying to improve the performance where the classifier $\mathbf{C}(x_i, \boldsymbol{A}, \mathbf{M})$ is commonly constructed as a function taking the embedded features and attributes to return the predicted confidence scores of all the classes represented in $\boldsymbol{A}$. We will review this family of methods according to their mainly relied frameworks.

Space alignment framework. These encoding based methods often have a specific embedding target space, which can be a commonly-used visual feature space, an manually defined semantic description space, or an unknown hidden space for detecting certain correlations. This idea is the first as well as one most common solution to zero-shot learning.

The classifier can be designed based on a fixed distance metric $\mathbf{d}(\cdot, \cdot)$ such as Euclidean distance or Cosine distance. Thereby, the predicted label for each visual feature $\mathbf{F}_f(x_i)$ is obtained as

$$\hat{y}_i = \arg\min_k (\mathbf{d}\left(\theta\left(\mathbf{F}_f(x_i)\right), \phi(a_k)\right), \quad s.t. \quad a_k \in \boldsymbol{A}_{te}. \tag{3.1}$$

In the following, we briefly review some representative work in the space alignment framework. In [113, 121], semantic-to-visual mappings are learned to align semantic and visual features from the same class. Specifically, the method in [121] utilizes a multi-layer neural network as the embedding function implying that the visual feature space is more appropriate as a target space to avoid aggravating the hubness problem, and a self-focus ratio according to the position of the embedded attributes is learned as an attention for each dimension of the visual feature space during optimization in [113]. More studies adopt the reconstruction or bi-direction mapping (a relaxed form of reconstruction) process to align the information from different spaces. Linear embedding functions are applied for both visual-to-semantic and semantic-to-visual projections in [41], and a rank minimization technique is additionally adopted for optimizing the linear transformation matrices. In [30], the encoding processes of the reconstruction are designed in both visual and semantic spaces, and achieve the joint embedding by minimizing the maximum mean discrepancy in the hidden layer. Then as a more strict case, the embeddings for the visual feature and semantic attributes from the same class are enforced to be equal in [118], and a two-alternate-steps algorithm is proposed in [53] to solve transformation matrices in the joint embedding with reconstruction supervision in two alternate steps. Similar classes for each class are selected via a threshold among cosine similarity in [4], then a semantic-to-visual-to-semantic reconstruction process is proposed, where the inter-class distances are pushed and the intra-class distances are reduced on the visual space. A projecting codebook is learned in [48] with an additional center loss in [24] and a reconstruction loss in [41] to embedded visual features and semantic attributes to a hidden orthogonal semantic space. The label space is selected as the embedding target space in [56], where the embedding of the unseen semantic attributes to the label space can be achieved by learning the projecting function from both the semantic and visual spaces to the label space. Such embedding is equivalent to linearly representing the labels of unseen classes by those of seen classes, thus improving the generalization of the model in the label space.

The classifier can also be designed learnable such as a bilinear function $\boldsymbol{W}$, which predicts the confidence scores as

$$\mathbf{C}\left(x_i, \boldsymbol{A}, \mathbf{M}, \boldsymbol{W}\right) = \theta\left(\mathbf{F}_f\left(x_i\right)\right)^T \boldsymbol{W}\,\phi\left(A\right). \tag{3.2}$$

The semantic attributes of both the seen and unseen classes are purely represented by those of seen in [122] to train the bilinear function which thus associates unseen classes with seen classes. Norms of the embedded semantic attributes and embedded visual feature is constrained in [77] for fair comparisons over classes and bounding the variation in semantic space respectively. In [120], the bilinear function is decomposed into two transformation matrices, and it is proved that minimizing the mean squared error between similarity matrices and the predicted scores for all samples is equivalent to restricting those transformation matrices to be orthogonal. A pairwise ranking loss function similar to the one in [102] is proposed in [17] as

$$\sum_{j \neq y_i}^{K^S} \left[I(j = y_i) + \mathbf{C}\left(x_i, a_j, \mathbf{M}, \boldsymbol{W}\right) - \mathbf{C}\left(x_i, a_{y_i}, \mathbf{M}, \boldsymbol{W}\right) \right]_+ . \tag{3.3}$$

Instead of the sum of all these pairwise terms, the ranking loss is modified by focusing on the pair. This leads to the maximum value in [1] and results in a weighted approximate one in [3] inspired by the unregularized ranking support vector machine [37]. It can also be redesigned with a triplet mining strategy to construct the triplet loss with the most negative samples and the most negative attributes as proposed in [34].

Moreover, the classifier can be defined in other forms. The instances from each class are assumed to follow an exponential family distribution in [93] where the parameters are learned from the semantic attributes. The method in [103] develops the ranking loss into a non-linear classifier case by learning multi-bilinear classifiers where each time this model chooses the one with the highest confidence score to be optimized. In [36] the attributes of unseen classes are utilized to reconstruct those of seen classes by the sparse coding approach. The solved coefficients are regarded as the similarity between classes. Then a neural network is designed to learn the similarity between the embedded attributes and visual features under the supervision of the labels and the similarities.

Graph based framework. A graph containing correlations between classes can be additionally constructed to enhance the generalization of the trained model. In [60], two relation graphs of the features in the hidden space are constructed based on the k-nearest neighbors among samples and the class labels which contribute to reducing distances between highly relevant features. This design is improved in [101] where two separated latent spaces are learned for embedding the visual samples and semantic attributes, and the k-nearest neighbor is replaced by the Cosine similarity to imply the relations among samples. Based on the two embedding spaces and the weighted sum of relations among samples and class labels an asymmetric graph structure with orthogonal projection is introduced to improve the learned latent space. By fixing the number of super-classes in different class layers, clusters obtained through the clustering algorithm among the attributes are taken to represent the super-class in [47], thereby a hierarchical graph over classes can be constructed to overcome the domain gap between seen and unseen classes. In [110], the relations among the classes are captured by augmenting the original label matrix in a dependency propagation process with the support of the low-rank constraint.

The graphic convolutional neural network (GCN) is a neural network that directly approximates localized spectral filters on graphs to learn hidden layer representations more relevant to the target task [40]. GCN is applied on the word embeddings of all the classes in [100] to learn the classifier parameters for each class. Then a dense graph propagation module is proposed in [38] where the connections from nodes to their ancestors and descendants are considered. In addition to the graph among word embeddings, in [98], the graph constructed through the k-nearest neighbor in the attribute space is also employed to learn the classifier parameters. The outputs of the GCN based on two graphs are weighted summed to learn the final parameters.

Meta learning framework. Meta learning process proposed in the few-shot learning aims to train models with high knowledge transfer ability [75]. In zero-shot learning, models trained on seen class data tend to overfit and perform poor on unseen classes. Therefore, the methods with similar meta learning strategies are developed to train more generalized models.

Relation network (RN) [88] is designed to learn a similarity measure based on the neural network architecture. The visual feature and embedded semantic attributes will be concatenated and used as the input to the measure model to return the similarity. The whole model is trained under a meta learning process where each time the loss function is designed based on a meta learning task sampled from the training set. Specifically, each time a small group of the samples are selected to construct the meta classification task where the number of the included classes is not settled. By training over several meta tasks, the trained model would be more adaptive for different tasks. Therefore, the model would be more generalized.

As an improvement of RN, CRnet [119] follows the same training process with the meta tasks. Additionally, an unsupervised K-means clustering algorithm is implemented to find the similar class groups and the corresponding group centers. Instead of training one embedding function among the semantic attributes, multi-attributes embedding functions are trained based on the group centers where the inputs are the differences between these centers and the semantic attributes. Then the sum of these embedded attributes is utilized for learning the similarity in the same way as RN.

A similar process is adopted in a correction network [27]. Based on the sampled meta tasks, an additional correction module is trained to modify the predicted value of the original model to become more precise. Then the learned correction module would be generalized since it is adapted to different meta tasks. As such, the correction will contribute to better performance.

3.2. Image-based methods

In the image-based methods, it is the original images $\boldsymbol{X}$ instead of the extracted feature vectors $\mathbf{F}_f(\boldsymbol{X})$ that are regarded as samples. Moreover, the well-designed backbone architecture with pre-trained parameters from the image classification task is partially or entirely borrowed as a learnable one $\mathbf{F} = \mathbf{F}_l$. The insight underlying these methods is to optimize the feature extractor $\mathbf{F}_l$ simultaneously with the specific designed embedding function and classifier. Sometimes an additional module accompanying the backbone is designed to obtain a more adaptable feature space, thus improving the performance.

Supervision based methods. By providing additional constraints or regularizations in the loss function for training, the feature extractor can be pushed to capture more relevant information, which results in a more representative feature space. Rather than training an embedding model with a bilinear classifier purely on the information from seen classes, unlabelled data are also employed in quasi-fully supervised learning [87]. Without supervised information, the predicted scores of the unseen classes for those unlabelled data are constrained to be large by constructing the sum of negative log values of them as a regularization term during optimization. Then training the whole model under this quasi-fully supervised setting with the designed loss will also improve the features extracted by the backbone. This can alleviate the bias towards seen classes.

A discriminative feature learning process is introduced in [51]. A zoomed coordinate is learned based on the feature maps to reconstruct a zoomed image sample with the same size as the original one, where visual features are extracted from both of the zoomed and original image samples. Since the semantic attributes are not discriminative enough, only a partial list of learned embedded features is adopted for learning the bilinear classifier with the attributes. Additionally, a triplet loss based on the squared Euclidean distance is constructed among the rest of the embedded features to improve the learned feature space.

Domain-aware visual bias eliminating [65] adopts a margin second-order embedding based on bilinear pooling [52] and a softmax loss function with a kind of temperature during training. As a result, the learned feature space constrained to be more discriminative leads to a low entropy for the instances from seen class. Then the instances from unseen class during the test would be distinguished with a relatively high entropy.

Attention based methods. As the attention mechanism has achieved significant performance in the image classification tasks [97], several attention relevant modules are also designed in zero-shot learning for capturing more representative features corresponding to the semantic information. In most of these methods, the attention module is utilized to obtain local features corresponding to certain specific semantic property. To produce more adequate supervision on the attention based feature space, a second-order operation [52] is applied on the learned features and semantics [108]. In the region graph embedding network [109], a transformation matrix is solved to represent the similarity between the attributes of the seen and the unseen classes. According to these similarities, a cross-entropy loss is then designed to ensure that the classifier also outputs a higher score for similar unseen classes when classifying samples from seen classes. As a result, the feature extractor is pushed to learn the feature space capturing more correlation information between seen and unseen classes. In [125], a triplet loss is designed to push the inter-class distances and reduce intra-class distances between features corresponding to both local and entire images. This model thus improves the learned feature space more conducive to the classification task.

Instead of purely training the attention module through the loss function defined on the feature space, additional explicit human annotated labels for attention can also be provided to supply the training. For example, in [58], captured gaze points are employed to generate the ground truth of the attention maps for constructing the binary cross-entropy loss across all the pixels. In addition to capturing local features, the attention learned from several feature maps is combined to guide the learning of the bilinear classifier [57].

3.3. Mechanism-improved methods

The insight of the mechanism-improved methods is to propose a generalized mechanism without changing or slightly changing the structure of the original method. The proposed mechanism can be an improvement of the training process, an optimization of a specific loss function, or a redesign prediction process. Commonly, this family of methods are designed for those zero-shot models sharing certain commonalities.

Training process focused. A theoretical explanation to normalization on attributes is presented in [83]. Then a more efficient normalization scheme is proposed standardizing the embedded attributes to alleviate the irregular loss surface.

During the feature extracting process, a fine-tuned backbone is proposed in the attribute prototype network (APN) [112]. In this work, assume the size of attributes is D_a. The prototype for each attribute $P = \{p_{d_a} \in \mathbb{R}^C\}_{d_a=1}^{D_a}$ is learned to generate similarity map $M^{d_a} = \{m_{i,j}^{d_a}\}^{h\times w}$ with height h and width w through multiplication of these prototypes and the corresponding feature maps. During the fine-tuning, the commonly used linear embedding classification loss is optimized with several regularization terms. An attribute decorrelation term is defined as the sum of l_2-norm of each dimension of the prototypes in the same disjoint attribute groups. This thus helps decorrelate unrelated attributes via enforcing prototypes in the same group sharing the value. Another similarity map compactness term can enforce the similarity maps concentrating on the peak region [123], which is given as

$$\mathcal{L}_{CPT} = \sum_{d_a=1}^{D_a} \sum_{i=1}^{h} \sum_{j=1}^{w} m_{i,j}^{d_a}[(i - \tilde{i})^2 + (j - \tilde{j})^2], \tag{3.4}$$

where $(\tilde{i}, \tilde{j})$ is the coordinate of the maximum value in M^{d_a}. This element-wise multiplication between the similarity map and the distance among coordinates constrains the similarity map to focus on a small number of local features. Thereby, each similarity map M^{d_a} can be regarded as the attention map corresponding to d_a-th attribute. The comparison result in this work shows that the fine-tuned backbone in APN outperforms the ones in some other methods [106, 124], even when fine-tuning is also implemented. In this sense, it can be regarded as a general improved one for feature extracting.

Isometric propagation network (IPN) [54] is proposed to guarantee the relation between classes in a propagation process based on a specific similarity measure. By defining the average of samples from the same class as the initialized visual class prototype, in the propagation, each time the prototype is re-represented by the weighted sum of the prototypes of similar classes. The similar classes are detected through a threshold and similarity measure which is the softmax with temperature among the Cosine similarity for each prototype. The similarity is also utilized as the weight for the re-representation. Such a propagation process can also be implemented on the semantic prototypes learned based on the trained semantic embedding module in other methods such as that used in [119]. During the test, the unseen prototypes could be obtained using the weighted sum of the propagated prototypes of seen classes according to the similarity measure, which contributes to significant performance improvement with the commonly used linear classification model.

The image is divided into different regions for extracting more precise features with the attention module in [31, 33, 82]. Moreover, an additional seen-unseen trade-off loss can be adopted to balance the predicted scores for seen and unseen classes. For example, a self-calibration loss term as a biased cross-entropy loss for the predicted unseen scores among samples from seen classes is designed in [31], and a soft cross-entropy loss based on the similarity between seen and unseen classes is utilized in [82]. Training the models with these additional constraints increases the prediction scores for unseen classes, thereby promoting the sensitivity of unseen class recognition.

A meta learning process with constructed meta training tasks is adopted in [75, 94] for few-shot learning. Instead of employing a loss function associated with the original classification task among the whole training set, several semi-tasks of the original task, namely meta tasks, are constructed with the meta training data sampled from the original training set. Adopting this meta learning process in zero-shot learning improves the generalization and restrains over-fitting [54, 88, 114, 119]. Figure 4 demonstrates an example of the meta zero-shot task in [88].

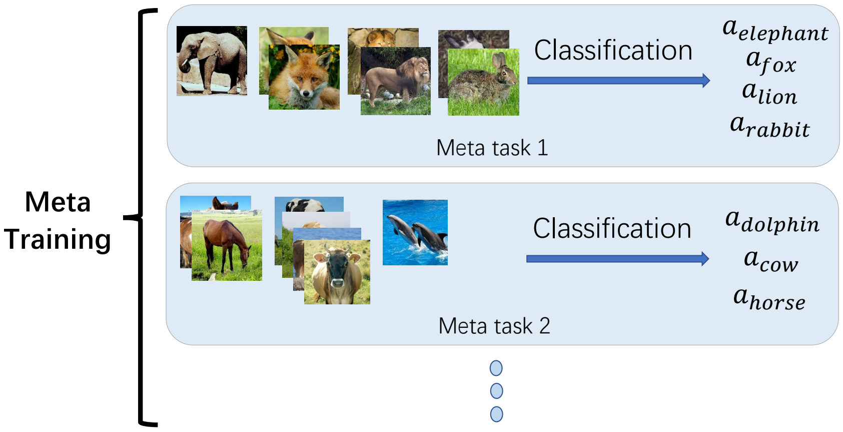

Figure 4. One illustrative example of meta tasks in meta learning process adopted in RN [88].

Test process focused. Since most of the methods suffer the class-lever over-fitting in generalized zero-shot tasks, a mechanism named Calibrated stacking is proposed in [9] to adjust the predicted confidence score for each class. With a trained classifier $\mathbf{C}$ and corresponding information extractor $\mathbf{M}$, the predicted confidence score in the regular test process can be obtained as $\mathbf{C}(x_i, \boldsymbol{A}_{te}, \mathbf{M})$. Then the prediction based on the calibrated stacking is defined as

$$\hat{y}_i = \arg\max_k \mathbf{C}(x_i, \boldsymbol{A}_{te}, \mathbf{M}) - \gamma I(k \leq K^S), \quad s.t. \quad a_k \in \boldsymbol{A}_{te}, \tag{3.5}$$

where $I()$ is the indicator function judging whether the k-th class belongs to a seen class and γ is the hyper-parameter controlling the scale of the adjustment. This calibrated stacking mechanism is simply subtracting a certain value for all the predicted seen confidence scores. Specifically, assume all the confidence scores are scaled in the range (0,1). Setting $\gamma = 1$ will lead that all the predicted labels belong to unseen classes, and conversely $\gamma = -1$ will cause all the predicted labels as seen classes. In other words, setting $\gamma = -1$ and 1 lead to zero accuracies for the unseen classes and seen classes, respectively. By adjusting γ from -1 to 1 with a tiny step size, one can obtain the adjusted accuracies for both the seen and unseen classes. Then a seen versus unseen accuracy curve can be plotted. In this case, the area under seen-unseen accuracy curve (AUSUC) is proposed as one optimal criterion measuring the overall performance of the models in generalized zero-shot learning tasks. A schematic is shown in Figure 5.

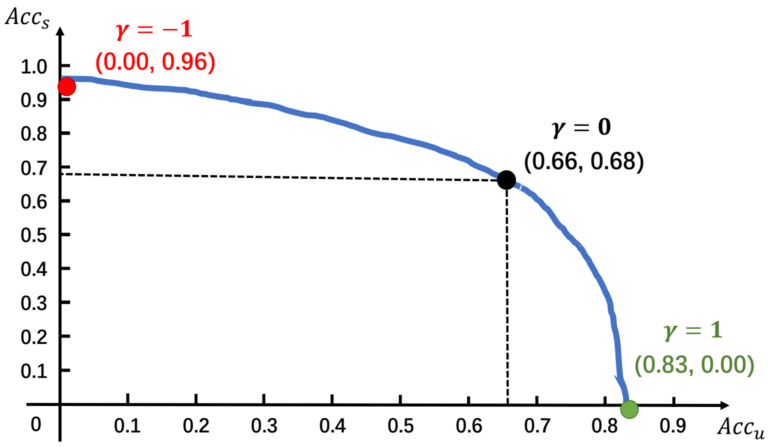

Figure 5. Schematic of the seen-unseen accuracy curve as defined in [9]. The black point with $\gamma = 0$ denotes the original performance of the model, the red point with $\gamma = -1$ and the green point with $\gamma = 1$ represent the adjusted results where the predicted scores are fully biased towards the seen and unseen classes, respectively.

Entire process focused. Instead of adjusting directly the confidence scores, a gradient based instance perturbation is introduced in [114]. A regularization term in [63] as the sum of the l_2-norm of input samples is adopted to achieve robust learning [21]. This training process could be regarded as adversarial defense which makes the learned classifier sufficiently robust to small perturbations in the sample space. During the test, the perturbed instance inclined to the unseen the most is obtained in the neighborhood of the original sample through calculating an adversarial perturbation based on a designed classification loss. Since the classifier is robust among the training classes, the predictions of unseen class instances will tend to be unseen, while those of seen class instances will keep consistent.

In [7], a self-learning process is proposed where each time hard unseen classes are selected based on the frequencies of the prediction during the test. Then an expanded training set with additional sampled instances from those hard unseen classes is constructed to re-train the model. The modified training set could enhance the sensitivity of model for those hard classes thus can boost the performance of the model under the transductive scenario.

4. Generative methods

The core component of generative methods is the generator that takes semantic information as input and outputs corresponding pseudo samples. Such a generator can be constructed based on variational autoencoder (VAE) [39] or generative adversarial network (GAN) [20] architecture. It can be also trained with the labelled samples with corresponded semantics. Then, by employing the unseen semantics, pseudo samples of unseen classes could be generated where the zero-shot learning task is converted to common classification. In this case, the information extractor $\mathbf{M}$ denotes a training process and the output is a trained generator $\mathbf{G}$ which takes $\boldsymbol{A}$ (sometimes combined with $\boldsymbol{X}_{tr}$) as inputs and outputs synthesized samples for corresponding classes. With the synthesized samples of unseen classes to support the training, the classifier can be designed as a common image classifier $C(\cdot)$ which takes samples as input and outputs the confidence score for each class. Here we will review those representative generative methods in different frameworks.

4.1. VAE based

Variational autoencoder is designed to derive a recognition model in the form $q_\phi(z|x)$ to approximate the real intractable posterior $p_{theta}(z|x)$ with the objective function:

$$\mathcal{L}(\theta, \phi, x_i) = -D_{KL}\left(q_\phi(z|x_i)\,||p_\theta(z)\right) + \mathbb{E}_{q_\phi(z|x_i)}\left[\log p_\theta(x_i|z)\right], \tag{4.1}$$

where D_{KL} denotes the Kullback-Leibler distance, $q_\phi(z|x)$ is regarded as a probabilistic encoder, and $p_\theta(x|z)$ is regarded as a probabilistic decoder. As the most straightforward form of VAE, conditional VAE [86] is applied to zero-shot learning in [66] as shown in Figure 6, where the sample is concatenated with the corresponding attributes to learn the distribution parameters; the sampled random variables based on the learned parameters are again concatenated with the corresponding attributes to reconstruct the sample. The objective function can be simply redesigned as

$$\mathcal{L}(\theta, \phi, x_i, a_{y_i}) = -D_{KL}\left(q_\phi\left(z|x_i, a_{y_i}\right)||p_\theta\left(z|a_{y_i}\right)\right) + \mathbb{E}_{q_\phi(z|x_i)}\left[\log p_\theta\left(x_i|z, a_{y_i}\right)\right]. \tag{4.2}$$

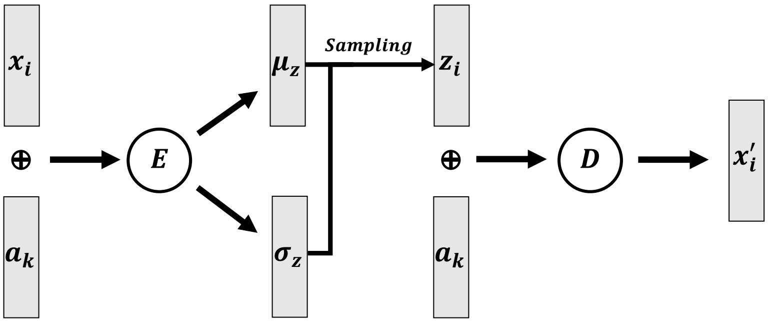

Figure 6. Schematic diagram of conditional VAE used in [66], where ⊕ denotes concatenation, E denotes the encoder, and D denotes the decoder.

In [90], Kullback-Leibler distance relevant to the synthesized samples and regression error of the semantic attributes from the corresponding synthesized samples are proposed as two additional regularization terms. A dual VAE architecture is designed in [81] where two VAE frameworks are trained respectively on the visual features and semantic attributes. The correlation between these two frameworks is constructed via minimizing the cross reconstruction errors and the Wasserstein distances between the latent Gaussian distribution for those sample-attributes pairs coming from the same class. The dual VAE is improved in [64], where a deep embedding network achieving the regression task from the semantic attribute to visual features is additionally designed. Then the hidden layer of this network is utilized as the input of the semantic VAE framework. The designed regression forces the hidden layer to become representative for both visual features and semantic attributes, thus benefiting the entire VAE framework. A disentangled dual VAE is designed in [50]. Different from the original dual VAE, each VAE framework learns two distributions, thereby sampling two random variables z_m^p and z_m^t. Notice that m denotes the modality which could be s and v representing semantic space and visual space respectively. For a group of pairs of training data, $\{z_{m,i}^p\}$ is shuffled as $\{\tilde{z}_{m,i}^p\}$ and then added up with $\{z_{m,i}^t\}$. Optimizing the model with this additional classification loss disentangles category-distilling factors and category-dispersing factors from both of the visual and semantic features. The multimodal VAE proposed in [5] builds one VAE framework for the concatenation of the visual feature and the embedded semantic attributes from the same class to capture the correlations between modalities. In identifiable VAE designed in [22], three VAE frameworks sharing the decoder for sample reconstruction are built taking the sample, the attribute, and both of them as inputs respectively. With an additional regularization term [42] encouraging disentanglement during inference, the learned latent space captures more significant information for generating discriminative samples.

4.2. GAN based

In generative adversarial networks, a generator G and a discriminator D are designed to be trained against each other iteratively with the loss function:

$$\min_G \max_D \mathcal{L}(D, G) = \mathbb{E}_{x \sim X_{tr}}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \tag{4.3}$$

Here, $p_z(z)$ denotes a prior on input noise variables z, the discriminator is trained to distinguish the generated pseudo samples from the samples in the original dataset and the target of generator is to synthesize pseudo samples as similar as the real samples so that the learned discriminator cannot recognize them. Then following the WGAN proposed in [23] where Wasserstein distance is leveraged, the loss of the conditional WGAN in zero-shot learning can be developed as

$$\begin{aligned}\min_G \max_D \mathcal{L}_{f-WGAN}(D, G) =& \mathbb{E}_{x \sim X_{tr}}\left[\log D\left(x, a_y\right)\right] - \mathbb{E}_{z \sim p_z(z)}\left[D\left(G\left(z, a_y\right), a_y\right)\right] \\ &- \lambda \mathbb{E}_{z \sim p_z(z)}\left[\left(\left\|\nabla_{G\left(z, a_y\right)} D\left(G\left(z, a_y\right), a_y\right)\right\|_2 - 1\right)^2\right].\end{aligned} \tag{4.4}$$

In [105], a classifier among seen classes is pre-trained on the training set, then adopted to supply a classification supervision for the samples generated from a WGAN framework. Guided by this additional supervision, the generator will learn to synthesize more discriminate samples which benefits the training of the final classifier. Inspired by the prototypical networks in few-shot learning [84], multiple prototypes of each seen class are calculated in [49]. Samples of each class are grouped into several clusters, then the average of samples in each group is regarded as one prototype for the corresponding class. Similarly, the prototypes of the synthesized samples could also be obtained based on the clusters. By minimizing the distances from the synthesized samples to their closest corresponding prototypes and distances from the synthesized prototypes to their closest real prototypes, the synthesized samples are constrained to be highly related to the attributes and real samples. Instead of adopting the classification supervision, a gradient guidance from a pre-trained classifier is proposed in [80]. In this model, classifier parameters at different spots during training are employed for calculating the optimization gradients based on the real sample and synthesized sample respectively. Expectations of the Cosine distances between the gradients are calculated from the real and synthesized samples, which are then utilized as an additional loss term to promote synthesizing samples as representative as real ones. In [25], conditional GAN is adopted with the designed instance-level and class-level contrastive embedding, where two classification problems are constructed among the embedded feature space to encourage the features to capture strong discriminative information. By employing additional taxonomy knowledge, hierarchical labels are obtained to calculate multi-prototypes for each class in [107]. Constraining the synthesized samples close to all their corresponding prototypes will encourage the synthesized samples to capture the hierarchical correlations. Inspired by space-aligned embedding, semantic rectifying GAN is proposed [117], in which a semantic rectifying loss is designed to enhance the discriminativeness of semantics under the guidance of visual relationships and two pre- and post-reconstructions (used to keep the consistency between synthesized visual and semantic features). Considering that the original semantics might not be discriminative enough, disentangling class representation generative adversarial network [116] is proposed to search automatically discriminative representations by a multi-modal triplet lossthat utilizes multi-modal information.

4.3. Muti-architecture based

Since GAN based methods tend to over-fit and VAE based methods tend to under-fit, some works adopt both the frameworks in their methods. CVAE is trained with a regressor against a discriminator in [28]. The framework proposed in [106] shares the decoder in conditional VAE as the generator for a conditional WGAN. This framework is also applicable for the transductive scenario by training another discriminator for unseen samples. In this model, a pre-trained classifier on the training set is adopted as classification supervision contributing to more discriminating synthesized samples. The dual VAE is trained with two additional discriminators in [62] based on the sum of the dual VAE loss and the conditional WGAN loss to avoid blurry synthesized samples.

4.4. Meta learning based

As a meta learning process proposed in [16], Model-Agnostic Meta-Learning is referred to in zero-shot learning to train generative models. First, each meta task contains meta training and meta validation set which are sampled from the training set. The model optimized over each meta task can become more generalized due to the divergence of the meta tasks. Moreover, the optimization process for parameters is also conducted in a meta way. Rather than learning parameters performing the best over tasks, the target here is to learn the most adaptive ones for all the meta tasks. In other words, the learned parameters may not achieve the best performance in the current training meta task, but may attain significant performance in different tasks with few-step training on them.

A conditional WGAN with a pre-trained classifier is optimized under this meta learning strategy in [91]. In [92], Model-Agnostic Meta-Learning is applied to the complex framework where the conditional VAE shares the decoder as the generator for a conditional WGAN. The parameters of the encoder, decoder (generator), and discriminator are optimized under such strategy to generate high-fidelity samples only relying on a few number of training examples from seen classes. Pseudo labels for the different meta task distribution is utilized for a task discriminator in [59]. During the training, once the task discriminator is defeated, the encoder is able to align multiple diverse tasks into a unified distribution. With the aligned embedded features, a conditional GAN which generates the pseudo embedded features from Gaussian noises and attributes with a learnable classifier can be trained under the meta learning strategy.

5. Implementation details

5.1. Benchmarks and evaluation criteria

Benchmarks. To avoid overlapping between unseen classes and training classes used for the pre-trained feature extractor, specific data splits for five commonly used benchmarks are proposed with extracted features in [104]. This work has greatly facilitated the evaluation of models for subsequent studies. Here, we will focus on four of them to set up a summary of the comparisons among the most representative methods.

Table 1. Statistics for AwA1, AwA2, aPY, CUB and SUN in terms of granularity, class size, sample size and sample divergence.

Dataset	Size	Granularity	Semantic type	Size of semantics	Class size		Sample size		
					train(seen)	unseen	train	test_{seen}	test_{unseen}
AwA1	medium	coarse	Attributes	85	40	10	19832	4958	5685
AwA2	medium	coarse	Attributes	85	40	10	23527	5882	7913
CUB	medium	fine	Attributes	312	150	50	7057	1764	2967
aPY	small	coarse	Attributes/text	64	20	12	5932	1483	7924
SUN	medium	fine	Attributes	102	645	72	10320	2580	1440

Animals with Attributes (AwA2) [104] contains 30,475 images from public web sources for 50 highly descriptive animal classes with at least 92 labelled examples per class. For example, the attributes include *stripes*, *brown*, *eats fish* and so on. Caltech-UCSD-Birds-200-2011 datasets (CUB)) [95] is a fine-grained dataset with a large number of classes and attributes, containing 11,788 images from 200 different types of birds annotated with 312 attributes. SUN Attribute (SUN) [72] is a fine-grained dataset, medium-scale in class number, containing 14,340 scene images annotated with 102 attributes, e.g. *sailing/boating*, *glass*, and *ocean*. The dataset Attribute Pascal and Yahoo (aPY) [15] is a small-scale dataset with 64 attributes and 32 object classes, including animals, vehicles, and buildings.

We recommend the splitting strategy used in [104] for the datasets, since most of the current methods are evaluated following such protocol. More details can be found in Table 1. Notice that Animals with Attributes (AwA1) [44] is not introduced here since it is not publicly available due to the copyright issue. It is worthy to mention that there are some other datasets adopted in zero-shot learning, e.g. the large scale dataset ImageNet-1K [12], the small scale fine-grained dataset Oxford Flower-102 (FLO) [68], and fMRI (functional Magnetic Resonance Images) [67]. Since they are not the most commonly used as the previous four benchmarks and some of the experimental settings on them are inconsistent in different studies, we will not go into details about them. Some evaluation protocols for them can be referred to in [8, 13, 17, 46, 70].

Evaluation criteria. Compared with the conventional zero-shot learning task, the generalized one can better evaluate the capability for constructing recognition conception of unseen classes, thus are selected for demonstrating the performances of the methods in this article. Since the model needs to discriminate between seen and unseen classes simultaneously ensuring correct classification, the performance of both seen and unseen classes needs to be measured. Following the most commonly used generalized task criteria defined in [104], we define ACC_S and ACC_U as two average per-class top-1 accuracies to measure the classification performances on seen and unseen classes as

$$ACC_S = \frac{1}{K^S}\sum_{k=1}^{K^S}\frac{TP_k}{N_k}, \tag{5.1}$$

$$ACC_U = \frac{1}{K^U}\sum_{k=1}^{K^U}\frac{TP_k}{N_k}, \tag{5.2}$$

where TP_k denotes the number of the true positive samples that is correctly predicted in kth-class and the N_k denotes the number of the instances in kth-class. In other words, the top-1 prediction accuracy for each class is considered equally independent of the sample size of that class. Specifically, the candidates for the predicted labels in such classification are all the classes but not singly those of seen or unseen. Then the comprehensive performance in generalized zero-shot learning task can be evaluated by the harmonic mean of these accuracies defined as follows:

$$H = \frac{2 \times ACC_S \times ACC_U}{ACC_S + ACC_U}. \tag{5.3}$$

5.2. Comparisons with implementation details

In this section, we will summarize the reported performance of the representative methods with implementation details. Tables 2 and 3 present the comparisons on the methods on AwA2, CUB, aPY, and SUN benchmarks in types of embedding methods and generative methods, respectively. The results are obtained from the corresponding published papers or the comparisons provided in [104] and all the H values are displayed in boldface. The listed methods are roughly sorted according to the published periods and performances for different scenarios. Here we regard the ResNet101 pre-trained on ImageNet 1K outputs features in 2,048 dimensions as the settled backbone for extracting visual features. In Table 2, the first part of the table divided by double solid lines presents the methods where the backbone is not changed and the rest summarizes those methods adjusting the backbone. The column Extra in the table contains several indicators about the implementation details that could boost the performance of the model, which are listed as follows.

- **Backbone modification.** Indicator $\mathbb{B}$ denotes that the architecture of the feature extractor is modified to improve the obtained visual feature space. Such modification includes designing additional attention modules accompanied with the backbone, repeatedly adopting the feature extractor to extract the divided image regions to obtain multiple features, employing the multi-channel feature map layer before pooling in the pre-trained ResNet, or constructing the backbone with other advanced neural network architectures.
- **Fine-tuning.** Indicator $\mathbb{F}$ specifies that the borrowed backbone is fine-tuned during training. As in most of the methods, the pre-trained backbone is frozen and the extracted visual features are directly employed as the training samples, their performances are evaluated under the same feature space. On the contrary, the methods fine-tuning the backbone with the proposed model lead to different feature spaces, thus the evaluation of them can not be considered strictly in the same setting as the methods without fine-tuning.

- **Additional knowledge.** Indicator $\mathbb{K}$ denotes that the information commonly not included in the benchmarks is leveraged to improve the performance of the model. Note that the pre-trained deep neural network is not counted as additional knowledge as this is somehow a common setting in zero-shot learning. Such additional knowledge includes taxonomy knowledge as hierarchical labels, correlations between attributes captured by manually defined or through word embedding models trained on extra text corpora, captured gaze point, and data augmentation technology.

Compared with the embedding methods of Table 2 in the same period, most of those generative methods of Table 3 appear to achieve better performance. Nonetheless, strictly speaking, training the classifier via samples generated based on unseen semantics in generative models can be considered as employing additionally unseen information (which is not used in embedding methods). Therefore, their performance difference may be due to such subtle setting difference. In this sense, to construct rigorous comparisons, we advocate evaluating the embedding and generative methods separately. Moreover, as the current best models in both of these two families under the inductive scenario, i.e. IPN [54] and CE-GZSL [25], perform quite similar actually, we believe embedding and generative methods are of equal importance in zero-shot learning.

Table 2. Comparisons of embedding methods on AwA2, CUB, aPY and SUN. Average ranking denotes the mean of the ranks of H values among the four datasets, "–" denotes the results were not reported, I, ST and T represent the inductive, semantic transductive, and transductive training scenarios respectively. Superscript with number denotes the same methods corresponding to different implementation setups.

Method	Scenario	Extra	AwA2			CUB			aPY			SUN			Average ranking
			ACC_U	ACC_S	**H**	ACC_U	ACC_S	**H**	ACC_U	ACC_S	**H**	ACC_U	ACC_S	**H**	
DeViSE (2013) [17]	I		17.1	74.7	**27.8**	23.8	53.0	**32.8**	4.9	76.9	**9.2**	16.9	27.4	**20.9**	13.8
SSE (2015) [122]	I		8.1	82.5	**14.8**	8.5	46.9	**14.4**	0.2	78.9	**0.4**	2.1	36.4	**4.0**	17.8
ESZSL (2015) [77]	I		5.9	77.8	**11.0**	12.6	63.8	**21.0**	2.4	70.1	**4.6**	11.0	27.9	**15.8**	17.0
SJE (2015) [3]	I		8.0	73.9	**14.4**	23.5	59.2	**33.6**	3.7	55.7	**6.9**	14.7	30.5	**19.8**	14.5
LatEm (2016) [103]	I		11.5	77.3	**20.0**	15.2	57.3	**24.0**	0.1	73.0	**0.2**	14.7	28.8	**19.5**	16.5
SAE (2017) [41]	I		1.1	82.2	**2.2**	7.8	54.0	**13.6**	0.4	80.9	**0.9**	8.8	18.0	**11.8**	18.3
DEM (2017) [121]	I		30.5	86.4	**45.1**	19.6	57.9	**29.2**	11.1	75.1	**19.4**	20.5	34.3	**25.6**	12.8
PSR (2018) [4]	I		20.7	73.8	**32.3**	24.6	54.3	**33.9**	13.5	51.4	**21.4**	20.8	37.2	**26.7**	11.5
LESAE (2018) [55]	I		21.8	70.6	**33.3**	24.3	53.0	**33.3**	12.7	56.1	**20.1**	21.9	34.7	**26.9**	11.8
RN (2018) [88]	I		30.0	93.4	**45.3**	38.1	61.1	**47.0**	–	–	–	–	–	–	9.5
SFDEM (2019) [113]	I		39.0	84.5	**53.4**	21.9	47.5	**30.0**	26.2	78.5	**39.3**	–	–	–	10.3
TVN (2019) [120]	I		–	–	–	26.5	62.3	**37.2**	16.1	66.9	**25.9**	22.2	38.3	**28.1**	9.7
PQZSL (2019) [48]	I		31.7	70.9	**43.8**	43.2	51.4	**46.9**	27.9	64.1	**38.8**	35.1	35.3	**35.2**	8.0
CRnet (2019) [119]	I		52.6	78.8	**63.1**	45.5	56.8	**50.5**	32.4	68.4	**44.0**	36.5	34.1	**35.3**	3.8
DTNet (2020) [34]	I		–	–	–	44.9	53.5	**48.9**	25.5	59.9	**35.5**	–	–	–	8.0
LAF (2020) [56]	I		50.4	58.5	**54.2**	43.7	52.0	**47.5**	33.8	49.0	**40.0**	36.0	36.6	**36.3**	5.5
advRN (2020) [114]	I		49.3	84.0	**62.2**	44.3	62.6	**51.9**	28.0	66.0	**39.3**	–	–	–	5.3
DVBE (2020)[1] [65]	I		63.6	70.8	**67.0**	53.2	60.2	**56.5**	32.6	58.3	**41.8**	45.0	37.2	**40.7**	2.5
LRSG-ZSL (2021) [110]	I		60.4	84.9	**70.6**	48.5	49.3	**48.9**	30.3	76.2	**43.4**	51.2	22.4	**31.2**	4.3
IPN (2021) [54]	I		67.5	79.2	**72.9**	60.2	73.8	**66.3**	37.2	66.0	**47.6**	–	–	–	1.0
TCN (2019) [36]	ST		61.2	65.8	**63.4**	52.6	52.0	**52.3**	24.1	64.0	**35.1**	31.2	37.3	**34.0**	5.5
LFGAA[1] (2019) [57]	I	$\mathbb{B}$ $\mathbb{F}$	27.0	93.4	**41.9**	36.2	80.9	**50.0**	–	–	–	18.5	40.0	**25.3**	11.0
AREN (2019) [108]	I	$\mathbb{B}$ $\mathbb{F}$	54.7	79.1	**64.7**	63.2	69.0	**66.0**	30.0	47.9	**36.9**	40.3	32.3	**35.9**	7.0
APN (2020) [112]	I	$\mathbb{B}$ $\mathbb{F}$ $\mathbb{K}$	56.5	78.0	**65.5**	65.3	69.3	**67.2**	–	–	–	41.9	34.0	**37.6**	6.3
DVBE (2020)[2] [65]	I	$\mathbb{F}$	62.7	77.5	**69.4**	64.4	73.2	**68.5**	37.9	55.9	**45.2**	44.1	41.6	**42.8**	3.5
GEM-ZSL (2021) [58]	I	$\mathbb{B}$ $\mathbb{F}$ $\mathbb{K}$	64.8	77.5	**70.6**	64.8	77.1	**70.4**	–	–	–	38.1	35.7	**36.9**	4.7
DAZLE (2020) [31]	ST	$\mathbb{B}$	60.3	75.7	**67.1**	56.7	59.6	**58.1**	–	–	–	52.3	24.3	**33.2**	8.3
RGEN (2020) [109]	ST	$\mathbb{B}$ $\mathbb{F}$	67.1	76.5	**71.5**	60.0	73.5	**66.1**	30.4	48.1	**37.2**	44.0	31.7	**36.8**	4.8
AGAN (2022) [82]	ST	$\mathbb{B}$	64.1	80.3	**71.3**	67.9	71.5	**69.7**	–	–	–	40.9	42.9	**41.8**	3.7
LFGAA[2] (2019) [57]	T	$\mathbb{B}$ $\mathbb{F}$	50.0	90.3	**64.4**	43.4	79.6	**56.2**	–	–	–	20.8	34.9	**26.1**	10.0
QFSL (2018) [87]	T	$\mathbb{F}$	66.2	93.1	**77.4**	71.5	74.9	**73.2**	–	–	–	51.3	31.2	**38.8**	2.3
STHS-S2V (2021) [7]	T		91.4	92.3	**91.8**	71.2	74.5	**72.8**	–	–	–	70.7	44.8	**54.8**	1.3

Table 3. Comparisons of generative methods on AwA2, CUB, aPY and SUN. Average ranking denotes the mean of the ranks of H values among the four datasets, "–" denotes the results were not reported, I, ST and T represent the inductive, semantic transductive, and transductive training scenarios respectively. Superscript with number denotes the same methods corresponding to different implementation setups.

Method	Scenario	Extra	AwA2			CUB			aPY			SUN			Average ranking
			ACC_U	ACC_S	H	ACC_U	ACC_S	H	ACC_U	ACC_S	H	ACC_U	ACC_S	H	
f-CLSWGAN (2018) [105]	I		–	–	–	43.7	57.7	**49.7**	–	–	–	42.6	36.6	**39.4**	18.0
SRGAN (2019) [117]	I		–	–	–	31.3	60.9	**41.3**	22.3	78.4	**34.8**	22.1	38.3	**27.4**	15.3
LisGAN (2019) [49]	I		–	–	–	46.5	57.9	**51.6**	–	–	–	42.9	37.8	**40.2**	17.0
GDAN (2019) [28]	I		32.1	67.5	**43.5**	39.3	66.7	**49.5**	30.4	75.0	**43.4**	38.1	89.9	**53.4**	11.0
CADA-VAE (2019) [81]	I		55.8	75.0	**63.9**	51.6	53.5	**52.4**	–	–	–	47.2	35.7	**40.6**	15.7
f-VAEGAN-D2^1 (2019) [106]	I		57.6	70.6	**63.5**	48.4	60.1	**53.6**	–	–	–	45.1	38.0	**41.3**	15.0
f-VAEGAN-D2^2 (2019) [106]	I	$\mathbb{F}$	57.1	76.1	**65.2**	63.2	75.6	**68.9**	–	–	–	50.1	37.8	**43.1**	9.3
ZSML (2020) [91]	I		58.9	74.6	**65.8**	60.0	52.1	**55.7**	36.3	46.6	**40.9**	–	–	–	10.0
DE-VAE (2020) [64]	I		58.8	78.9	**67.4**	52.5	56.3	**54.3**	–	–	–	45.9	36.9	**40.9**	12.0
DR-VAE (2021) [50]	I		56.9	80.2	**66.6**	51.1	58.2	**54.4**	–	–	–	36.6	47.6	**41.4**	11.7
M-VAE (2021) [5]	I		61.3	72.4	**66.4**	57.1	62.9	**59.8**	–	–	–	42.4	58.7	**49.2**	8.3
DGN (2021) [111]	I		60.1	76.4	**67.3**	53.8	61.9	**57.6**	36.5	61.7	**45.9**	48.3	37.4	**42.1**	8.5
DCRGAN (2021) [116]	I		–	–	–	55.8	66.8	**60.8**	37.2	71.7	**49.0**	47.1	38.5	**42.4**	6.3
CE-GZSL (2021) [25]	I		63.1	78.6	**70.0**	63.9	66.8	**65.3**	–	–	–	48.8	38.6	**43.1**	7.0
TGMZ (2021) [59]	I	$\mathbb{K}$	64.1	77.3	**70.1**	60.3	56.8	**58.5**	34.8	77.1	**48.0**	–	–	–	6.0
CKL+TR (2021) [107]	I	$\mathbb{K}$	61.2	92.6	**73.7**	57.8	50.2	**53.7**	30.8	78.9	**44.3**	–	–	–	8.0
APN+f-VAEGAN-D2 (2020) [112]	I	$\mathbb{B}\,\mathbb{F}\,\mathbb{K}$	62.2	69.5	**65.6**	65.7	74.9	**70.0**	–	–	–	49.4	39.2	**43.7**	8.0
AFGN (2022) [82]	ST	$\mathbb{B}$	68.1	82.9	**74.7**	69.8	77.1	**73.2**	–	–	–	53.1	45.9	**49.2**	3.7
f-VAEGAN-D2^3 (2019) [106]	T		84.8	88.6	**86.7**	61.4	65.1	**63.2**	–	–	–	60.6	41.9	**49.6**	4.3
f-VAEGAN-D2^4 (2019) [106]	T	$\mathbb{F}$	86.3	88.7	**87.5**	73.8	81.4	**77.3**	–	–	–	54.2	41.8	**47.2**	3.0
STHS-WGAN (2021) [7]	T		94.9	92.3	**93.6**	77.4	74.5	**75.9**	–	–	–	67.5	44.8	**53.9**	1.3

Moreover, as shown in these two tables, the methods with modified or fine-tuned backbones outperform their original counterparts published in the same year. Especially, the effectiveness of fine-tuning has been verified in the embedding method DVBE [65] and the generative method f-VAEGAN-D2 [106]. Fine-tuning leads to 2.4%, 12.0%, 3.4%, 2.1% absolute increment in the H values for DVBE on AwA2, CUB, aPY and SUN respectively. Similar improvements can also be observed for the f-VAEGAN-D2 under the inductive and transductive scenarios. These results imply that fine-tuning the backbone overall benefits the generalized zero-shot learning especially on the CUB benchmarks.

Most outstanding embedding and generative methods under the inductive scenario often utilize additional knowledge. In this way, more adequate information can help better construct concepts of unseen classes through knowledge of seen classes. The validity of the employed additional knowledge is not accurately presented in these comparison tables. One can refer to each relevant paper for more details.

When the methods in all the scenarios are compared together, for both the embedding and generative methods, one can find that methods STHS-S2V and STHS-WGAN [7] in the transductive scenario, attain the highest H values on most of the benchmarks. The unlabelled data with unseen classes attributes provide a detailed target guidance for the transformation of categorical knowledge, thus making such scenario the easiest generalized case. If one takes TCN [36] as the most similar method of RN [88] under the semantic transductive scenario (via accessing the unseen attributes during training), 18.1% and 5.3% absolute improvement have been achieved on AwA2 and CUB, respectively. Moreover, the gaps between the performances of the LFGAA [57] under both the semantic transductive and inductive scenarios also confirm the contribution of unseen attributes in training the model in generalized zero-shot learning.

In this section, the type of the classifiers or the number of synthesized pseudo samples for training is not collated here, as the impact of these implementation details on model performance is uncertain when the models are structured differently or applied on different databases. We focus on specifying differences in the implementation details which commonly lead to explicit changes in performance among the current representative methods. On the one hand, we acknowledge the contribution of those methods of adopting additional knowledge or modifications; on the other hand, showcasing numerical comparisons among different methods with different implementation settings may not be rigorous enough, which could lead to a misleading assessment of the capability of the model. We advocate researchers set up comparisons among the methods under the same implementation settings. Moreover, all the additional operations and/or auxiliary knowledge appear critically important and thus should keep clear and be stated explicitly for fair and precise evaluations.

6. Conclusions

In this article, we have provided a comprehensive survey of image classification with zero-shot learning. We have put one main focus of this survey on the implementation issues. Particularly, with the methods steadily improved, different problem settings, and diverse experimental setups have emerged, and thus we have examined three implementation details that can boost the performance of zero-shot learning, i.e. whether the backbone structure has been modified, whether fine-tuning has been conducted, and whether additional knowledge has been used. By annotating these experimental details, we have collected a more careful comparison among various zero-shot methodologies. While generative methods appear to outperform embedding methods overall, we argue that the performance difference may be due to the different settings, thus suggesting that it may be fairer to compare them separately. Moreover, we observe that the current best models in both families perform quite similar under the inductive scenario. Thus we believe embedding and generative methods are of equal importance in zero-shot learning.

Acknowledgments

The work was partially supported by the following: National Natural Science Foundation of China under no.61876155; Jiangsu Science and Technology Programme (Natural Science Foundation of Jiangsu Province) under no. BE2020006-4; Key Program Special Fund in XJTLU under no. KSF-T-06, and KSF-E-26.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for attribute-based classification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2013), 819–826. https://doi.org/10.1109/CVPR.2013.111
2. Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for image classification, *IEEE T. Pattern Anal.*, **38** (2015), 1425–1438. https://doi.org/10.1109/TPAMI.2015.2487986

3. Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele, Evaluation of output embeddings for fine-grained image classification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 2927–2936. https://doi.org/10.1109/CVPR.2015.7298911
4. Y. Annadani, S. Biswas, Preserving semantic relations for zero-shot learning, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 7603–7612.
5. N. Bendre, K. Desai, P. Najafirad, Generalized zero-shot learning using multimodal variational auto-encoder with semantic concepts, *IEEE International Conference on Image Processing (ICIP)*, (2021), 1284–1288. https://doi.org/10.1109/ICIP42928.2021.9506108
6. I. Biederman, Recognition-by-components: a theory of human image understanding, *Psychol. Rev.*, **94** (1987), 115. 10.1037/0033-295X.94.2.115
7. L. Bo, Q. Dong, Z. Hu, Hardness sampling for self-training based transductive zero-shot learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 16499–16508. https://doi.org/10.1109/CVPR46437.2021.01623
8. S. Changpinyo, W.-L. Chao, B. Gong, F. Sha, Synthesized classifiers for zero-shot learning, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 5327–5336. https://doi.org/10.1109/CVPR.2016.575
9. W.-L. Chao, S. Changpinyo, B. Gong, F. Sha, An empirical study and analysis of generalized zero-shot learning for object recognition in the wild, *European Conference on Computer Vision (ECCV)*, (2016), 52–68. https://doi.org/10.1007/978-3-319-46475-6_4
10. Q. Chen, W. Wang, K. Huang, F. Coenen, Zero-shot text classification via knowledge graph embedding for social media data, *IEEE Internet Things*, (2021). https://doi.org/10.1109/JIOT.2021.3093065
11. Y.-J. Chen, Y.-J. Chang, S.-C. Wen, Y. Shi, X. Xu, T.-Y. Ho, Q. Jia, M. Huang, et al., Zero-shot medical image artifact reduction, *IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, (2020), 862–866. https://doi.org/10.1109/ISBI45749.2020.9098566
12. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2009), 248–255. https://doi.org/10.1109/CVPR.2009.5206848
13. M. Elhoseiny, B. Saleh, A. Elgammal, Write a classifier: Zero-shot learning using purely textual descriptions, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2013), 2584–2591. https://doi.org/10.1109/ICCV.2013.321
14. M. Elhoseiny, Y. Zhu, H. Zhang, A. Elgammal, Link the head to the“ beak”: Zero shot learning from noisy text description at part precision, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 5640–5649. https://doi.org/10.1109/CVPR.2017.666
15. A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2009), 1778–1785. https://doi.org/10.1109/CVPR.2009.5206772
16. C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, *Proceedings of the 34th International Conference on Machine Learning (ICML)*, **70** (2017), 1126–1135.

17. A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, T. Mikolov, Devise: A deep visual-semantic embedding model, *Advances in neural information processing systems*, **26** (2013), 2121–2129.
18. Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, Transductive multi-view zero-shot learning, *IEEE T. Pattern Anal.*, **37** (2015), 2332–2345. https://doi.org/10.1109/TPAMI.2015.2408354
19. Z. Fu, T. Xiang, E. Kodirov, S. Gong, Zero-shot object recognition by semantic manifold distance, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 2635–2644. https://doi.org/10.1109/CVPR.2015.7298879
20. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, *Advances in neural information processing systems*, (2014), 2672–2680.
21. I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, *3rd International Conference on Learning Representations (ICLR)*, (2015).
22. M. Gull, O. Arif, Generalized zero-shot learning using identifiable variational autoencoders, *Expert Syst. Appl.*, **191** (2022), 116268. https://doi.org/10.1016/j.eswa.2021.116268
23. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, Improved training of wasserstein gans, *Advances in neural information processing systems*, **30** (2017), 5767–5777.
24. Y. Guo, G. Ding, J. Han, Y. Gao, Sitnet: Discrete similarity transfer network for zero-shot hashing., *the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, (2017), 1767–1773. https://doi.org/10.24963/ijcai.2017/245
25. Z. Han, Z. Fu, S. Chen, J. Yang, Contrastive embedding for generalized zero-shot learning, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 2371–2381. https://doi.org/10.1109/CVPR46437.2021.00240
26. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
27. R. L. Hu, C. Xiong, R. Socher, *Correction networks: Meta-learning for zero-shot learning*, 2018.
28. H. Huang, C. Wang, P. S. Yu, C.-D. Wang, Generative dual adversarial network for generalized zero-shot learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 801–810. https://doi.org/10.1109/CVPR.2019.00089
29. K. Huang, A. Hussain, Q.-F. Wang, R. Zhang, *Deep Learning: Fundamentals, Theory and Applications*, Springer, 2019.
30. Y.-H. Hubert Tsai, L.-K. Huang, R. Salakhutdinov, Learning robust visual-semantic embeddings, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2017), 3571–3580. https://doi.org/10.1109/ICCV.2017.386
31. D. Huynh, E. Elhamifar, Fine-grained generalized zero-shot learning via dense attribute-based attention, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 4483–4493. https://doi.org/10.1109/CVPR42600.2020.00454
32. D. Jayaraman, F. Sha, K. Grauman, Decorrelating semantic visual attributes by resisting the urge to share, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2014), 1629–1636. https://doi.org/10.1109/CVPR.2014.211

33. Z. Ji, Y. Fu, J. Guo, Y. Pang, Z. M. Zhang, Stacked semantics-guided attention model for fine-grained zero-shot learning, *Advances in Neural Information Processing Systems*, **31** (2018), 5998–6007.
34. Z. Ji, H. Wang, Y. Pang, L. Shao, Dual triplet network for image zero-shot learning, *Neurocomputing*, **373** (2020), 90–97. https://doi.org/10.1016/j.neucom.2019.09.062
35. H. Jiang, G. Yang, K. Huang, R. Zhang, W-net: one-shot arbitrary-style chinese character generation with deep neural networks, *International Conference on Neural Information Processing*, (2018), 483–493. https://doi.org/10.1007/978-3-030-04221-9_43
36. H. Jiang, R. Wang, S. Shan, X. Chen, Transferable contrastive network for generalized zero-shot learning, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2019), 9765–9774. https://doi.org/10.1109/ICCV.2019.00986
37. T. Joachims, Optimizing search engines using clickthrough data, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, (2002), 133–142. https://doi.org/10.1145/775047.775067
38. M. Kampffmeyer, Y. Chen, X. Liang, H. Wang, Y. Zhang, E. P. Xing, Rethinking knowledge graph propagation for zero-shot learning, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 11487–11496. https://doi.org/10.1109/CVPR.2019.01175
39. D. P. Kingma, M. Welling, Auto-encoding variational bayes, *2nd International Conference on Learning Representations (ICLR)*, (2014).
40. T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, *5th International Conference on Learning Representations (ICLR)*, OpenReview.net, (2017).
41. E. Kodirov, T. Xiang, S. Gong, Semantic autoencoder for zero-shot learning, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 3174–3183. https://doi.org/10.1109/CVPR.2017.473
42. A. Kumar, P. Sattigeri, A. Balakrishnan, Variational inference of disentangled latent concepts from unlabelled observations, *6th International Conference on Learning Representations (ICLR)*, (2018).
43. C. H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2009), 951–958. https://doi.org/10.1109/CVPR.2009.5206594
44. C. H. Lampert, H. Nickisch, S. Harmeling, Attribute-based classification for zero-shot visual object categorization, *IEEE T. Pattern Anal.*, **36** (2014), 453–465. https://doi.org/10.1109/TPAMI.2013.140
45. H. Larochelle, D. Erhan, Y. Bengio, Zero-data learning of new tasks, *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, (2008), 646–651.
46. J. Lei Ba, K. Swersky, S. Fidler, Predicting deep zero-shot convolutional neural networks using textual descriptions, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2015), 4247–4255. https://doi.org/10.1109/ICCV.2015.483
47. A. Li, Z. Lu, J. Guan, T. Xiang, L. Wang, J. Wen, Transferrable feature and projection learning with class hierarchy for zero-shot learning, *Int. J. Comput. Vis.*, **128** (2020), 2810–2827. https://doi.org/10.1007/s11263-020-01342-x

48. J. Li, X. Lan, Y. Liu, L. Wang, N. Zheng, Compressing unknown images with product quantizer for efficient zero-shot classification, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 5463–5472. https://doi.org/10.1109/CVPR.2019.00561
49. J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, Z. Huang, Leveraging the invariant side of generative zero-shot learning, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 7402–7411. https://doi.org/10.1109/CVPR.2019.00758
50. X. Li, Z. Xu, K. Wei, C. Deng, Generalized zero-shot learning via disentangled representation, *Proceedings of the AAAI Conference on Artificial Intelligence*, **35** (2021), 1966–1974.
51. Y. Li, J. Zhang, J. Zhang, K. Huang, Discriminative learning of latent features for zero-shot recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 7463–7471. https://doi.org/10.1109/CVPR.2018.00779
52. T.-Y. Lin, A. RoyChowdhury, S. Maji, Bilinear cnn models for fine-grained visual recognition, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2015), 1449–1457. https://doi.org/10.1109/ICCV.2015.170
53. G. Liu, J. Guan, M. Zhang, J. Zhang, Z. Wang, Z. Lu, Joint projection and subspace learning for zero-shot recognition, *2019 IEEE International Conference on Multimedia and Expo (ICME)*, (2019), 1228–1233. https://doi.org/10.1109/ICME.2019.00214
54. L. Liu, T. Zhou, G. Long, J. Jiang, X. Dong, C. Zhang, Isometric propagation network for generalized zero-shot learning, *9th International Conference on Learning Representations (ICLR)*, OpenReview.net, (2021).
55. Y. Liu, Q. Gao, J. Li, J. Han, L. Shao, Zero shot learning via low-rank embedded semantic autoencoder, *the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, (2018), 2490–2496. https://doi.org/10.24963/ijcai.2018/345
56. Y. Liu, X. Gao, Q. Gao, J. Han, L. Shao, Label-activating framework for zero-shot learning, *Neural Networks*, **121** (2020), 1–9. https://doi.org/10.1016/j.neunet.2019.08.023
57. Y. Liu, J. Guo, D. Cai, X. He, Attribute attention for semantic disambiguation in zero-shot learning, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 6698–6707. https://doi.org/10.1109/ICCV.2019.00680
58. Y. Liu, L. Zhou, X. Bai, Y. Huang, L. Gu, J. Zhou, T. Harada, Goal-oriented gaze estimation for zero-shot learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 3794–3803. https://doi.org/10.1109/CVPR46437.2021.00379
59. Z. Liu, Y. Li, L. Yao, X. Wang, G. Long, Task aligned generative meta-learning for zero-shot learning, *Proceedings of The Thirty-Fifth AAAI Conference on Artificial Intelligence*, (2021).
60. Y. Long, L. Liu, L. Shao, Attribute embedding with visual-semantic ambiguity removal for zero-shot learning, *Proceedings of the British Machine Vision Conference 2016*, BMVA Press, (2016). https://doi.org/10.5244/C.30.40
61. Y. Long, L. Liu, L. Shao, F. Shen, G. Ding, J. Han, From zero-shot learning to conventional supervised classification: Unseen visual data synthesis, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 1627–1636. https://doi.org/10.1109/CVPR.2017.653

62. Y. Luo, X. Wang, F. Pourpanah, Dual VAEGAN: A generative model for generalized zero-shot learning, *Appl. Soft Comput.*, **107** (2021), 107352. https://doi.org/10.1016/j.asoc.2021.107352
63. C. Lyu, K. Huang, H.-N. Liang, A unified gradient regularization family for adversarial examples, *2015 IEEE international conference on data mining*, (2015), 301–309. https://doi.org/10.1109/ICDM.2015.84
64. P. Ma, X. Hu, A variational autoencoder with deep embedding model for generalized zero-shot learning, *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2020), 11733–11740. https://doi.org/10.1609/aaai.v34i07.6844
65. S. Min, H. Yao, H. Xie, C. Wang, Z.-J. Zha, Y. Zhang, Domain-aware visual bias eliminating for generalized zero-shot learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 12664–12673. https://doi.org/10.1109/CVPR42600.2020.01268
66. A. Mishra, S. Krishna Reddy, A. Mittal, H. A. Murthy, A generative model for zero shot learning using conditional variational autoencoders, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) workshops*, (2018), 2188–2196. https://doi.org/10.1109/CVPRW.2018.00294
67. T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, M. A. Just, Predicting human brain activity associated with the meanings of nouns, *science*, **320** (2008), 1191–1195. https://doi.org/10.1126/science.1152876
68. M. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, (2008), 722–729. https://doi.org/10.1109/ICVGIP.2008.47
69. M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, J. Dean, Zero-shot learning by convex combination of semantic embeddings, *2nd International Conference on Learning Representations (ICLR)* (eds. Y. Bengio and Y. LeCun), (2014).
70. M. Palatucci, D. Pomerleau, G. E. Hinton, T. M. Mitchell, Zero-shot learning with semantic output codes, *Advances in neural information processing systems*, (2009), 1410–1418.
71. D. Parikh, K. Grauman, Relative attributes, *2011 International Conference on Computer Vision (ICCV)*, (2011), 503–510. https://doi.org/10.1109/ICCV.2011.6126281
72. G. Patterson, J. Hays, SUN attribute database: Discovering, annotating, and recognizing scene attributes, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2012), 2751–2758.
73. F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, A review of generalized zero-shot learning methods, *arXiv preprint arXiv:2011.08641*.
74. R. Qiao, L. Liu, C. Shen, A. Van Den Hengel, Less is more: zero-shot learning from online textual documents with noise suppression, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 2249–2257. https://doi.org/10.1109/CVPR.2016.247
75. S. Ravi, H. Larochelle, Optimization as a model for few-shot learning, *5th International Conference on Learning Representations (ICLR)*, OpenReview.net, (2017).
76. S. Reed, Z. Akata, H. Lee, B. Schiele, Learning deep representations of fine-grained visual descriptions, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 49–58. https://doi.org/10.1109/CVPR.2016.13

77. B. Romera-Paredes, P. Torr, An embarrassingly simple approach to zero-shot learning, *International Conference on Machine Learning (ICML)*, (2015), 2152–2161.

78. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, et al., Imagenet large scale visual recognition challenge, *Int. J. Comput. Vision*, **115** (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

79. C. Samplawski, E. Learned-Miller, H. Kwon, B. M. Marlin, Zero-shot learning in the presence of hierarchically coarsened labels, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, (2020), 926–927. https://doi.org/10.1109/CVPRW50498.2020.00471

80. M. B. Sariyildiz, R. G. Cinbis, Gradient matching generative networks for zero-shot learning, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 2168–2178. https://doi.org/10.1109/CVPR.2019.00227

81. E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, Z. Akata, Generalized zero-and few-shot learning via aligned variational autoencoders, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 8247–8255. https://doi.org/10.1109/CVPR.2019.00844

82. T. Shermin, S. W. Teng, F. Sohel, M. Murshed, G. Lu, Integrated generalized zero-shot learning for fine-grained classification, *Pattern Recogn.*, **122** (2022), 108246. https://doi.org/10.1016/j.patcog.2021.108246

83. I. Skorokhodov, M. Elhoseiny, Class normalization for (continual)? generalized zero-shot learning, *9th International Conference on Learning Representations (ICLR)*, OpenReview.net, 2021.

84. J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, *Advances in Neural Information Processing Systems*, (2017), 4077–4087.

85. R. Socher, M. Ganjoo, C. D. Manning, A. Y. Ng, Zero-shot learning through cross-modal transfer, *Advances in Neural Information Processing Systems*, (2013), 935–943.

86. K. Sohn, H. Lee and X. Yan, Learning structured output representation using deep conditional generative models, *Advances in neural information processing systems*, **28** (2015), 3483–3491.

87. J. Song, C. Shen, Y. Yang, Y. Liu, M. Song, Transductive unbiased embedding for zero-shot learning, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 1024–1033. https://doi.org/10.1109/CVPR.2018.00113

88. F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, T. M. Hospedales, Learning to compare: Relation network for few-shot learning, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 1199–1208. https://doi.org/10.1109/CVPR.2018.00131

89. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, et al., Going deeper with convolutions, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 1–9. https://doi.org/10.1109/CVPR.2015.7298594

90. V. K. Verma, G. Arora, A. Mishra, P. Rai, Generalized zero-shot learning via synthesized examples, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 4281–4289. https://doi.org/10.1109/CVPR.2018.00450

91. V. K. Verma, D. Brahma, P. Rai, Meta-learning for generalized zero-shot learning, *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, (2020), 6062–6069. https://doi.org/10.1609/aaai.v34i04.6069
92. V. K. Verma, A. Mishra, A. Pandey, H. A. Murthy, P. Rai, Towards zero-shot learning with fewer seen class examples, *IEEE Winter Conference on Applications of Computer Vision*, (2021), 2240–2250. https://doi.org/10.1109/WACV48630.2021.00229
93. V. K. Verma, P. Rai, A simple exponential family framework for zero-shot learning, *Joint European conference on machine learning and knowledge discovery in databases*, (2017), 792–808. https://doi.org/10.1007/978-3-319-71246-8_48
94. O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, Matching networks for one shot learning, *Advances in neural information processing systems*, **29** (2016), 3630–3638.
95. C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, *The caltech-ucsd birds-200-2011 dataset*, California Institute of Technology, 2011.
96. D. Wang, Y. Li, Y. Lin, Y. Zhuang, Relational knowledge transfer for zero-shot learning, *Thirtieth AAAI Conference on Artificial Intelligence*, (2016).
97. F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 3156–3164. https://doi.org/10.1109/CVPR.2017.683
98. J. Wang, B. Jiang, Zero-shot learning via contrastive learning on dual knowledge graphs, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 885–892. https://doi.org/10.1109/ICCVW54120.2021.00104
99. W. Wang, V. W. Zheng, H. Yu, C. Miao, A survey of zero-shot learning: Settings, methods, and applications, *ACM T. Intel. Syst. Tec. (TIST)*, **10** (2019), 1–37.
100. X. Wang, Y. Ye, A. Gupta, Zero-shot recognition via semantic embeddings and knowledge graphs, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 6857–6866. https://doi.org/10.1109/CVPR.2018.00717
101. Y. Wang, H. Zhang, Z. Zhang, Y. Long, Asymmetric graph based zero shot learning, *Multim. Tools Appl.*, **79** (2020), 33689–33710. https://doi.org/10.1007/s11042-019-7689-y
102. J. Weston, S. Bengio, N. Usunier, Wsabie: Scaling up to large vocabulary image annotation, *the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, (2011).
103. Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, B. Schiele, Latent embeddings for zero-shot classification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 69–77. https://doi.org/10.1109/CVPR.2016.15
104. Y. Xian, C. H. Lampert, B. Schiele, Z. Akata, Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly, *IEEE T. Pattern Anal.*, **41** (2018), 2251–2265.
105. Y. Xian, T. Lorenz, B. Schiele, Z. Akata, Feature generating networks for zero-shot learning, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 5542–5551. https://doi.org/10.1109/CVPR.2018.00581
106. Y. Xian, S. Sharma, B. Schiele, Z. Akata, f-vaegan-d2: A feature generating framework for any-shot learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 10275–10284. https://doi.org/10.1109/CVPR.2019.01052

107. C. Xie, H. Xiang, T. Zeng, Y. Yang, B. Yu, Q. Liu, Cross knowledge-based generative zero-shot learning approach with taxonomy regularization, *Neural Networks*, **139** (2021), 168–178. https://doi.org/10.1016/j.neunet.2021.02.009

108. G.-S. Xie, L. Liu, X. Jin, F. Zhu, Z. Zhang, J. Qin, Y. Yao, L. Shao, Attentive region embedding network for zero-shot learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 9384–9393. https://doi.org/10.1109/CVPR.2019.00961

109. G.-S. Xie, L. Liu, F. Zhu, F. Zhao, Z. Zhang, Y. Yao, J. Qin, L. Shao, Region graph embedding network for zero-shot learning, *European Conference on Computer Vision (ECCV)*, (2020), 562–580. https://doi.org/10.1007/978-3-030-58548-8_33

110. B. Xu, Z. Zeng, C. Lian, Z. Ding, Semi-supervised low-rank semantics grouping for zero-shot learning, *IEEE Trans. Image Process.*, **30** (2021), 2207–2219. https://doi.org/10.1109/TIP.2021.3050677

111. T. Xu, Y. Zhao, X. Liu, Dual generative network with discriminative information for generalized zero-shot learning, *Complexity*, **2021** (2021), 6656797:1–6656797:11. https://doi.org/10.1155/2021/6656797

112. W. Xu, Y. Xian, J. Wang, B. Schiele, Z. Akata, Attribute prototype network for zero-shot learning, *Advances in Neural Information Processing Systems*, **33** (2020), 21969–21980.

113. G. Yang, K. Huang, R. Zhang, J. Y. Goulermas, A. Hussain, Self-focus deep embedding model for coarse-grained zero-shot classification, *International Conference on Brain Inspired Cognitive Systems*, (2019), 12–22.

114. G. Yang, K. Huang, R. Zhang, J. Y. Goulermas, A. Hussain, Inductive generalized zero-shot learning with adversarial relation network, *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, (2020), 724–739.

115. B. Yao, A. Khosla, L. Fei-Fei, Combining randomization and discrimination for fine-grained image categorization, *CVPR 2011*, (2011), 1577–1584. https://doi.org/10.1109/CVPR.2011.5995368

116. Z. Ye, F. Hu, F. Lyu, L. Li, K. Huang, Disentangling semantic-to-visual confusion for zero-shot learning, *IEEE T. Multimedia*, (2021). doi:10.1109/TMM.2021.3089017. https://doi.org/10.1109/TMM.2021.3089017

117. Z. Ye, F. Lyu, L. Li, Q. Fu, J. Ren, F. Hu, Sr-gan: Semantic rectifying generative adversarial network for zero-shot learning, *2019 IEEE International Conference on Multimedia and Expo (ICME)*, (2019), 85–90. https://doi.org/10.1109/ICME.2019.00023

118. Y. Yu, Z. Ji, J. Guo, Z. Zhang, Zero-shot learning via latent space encoding, *IEEE T. Cybernetics*, **49** (2018), 3755–3766. https://doi.org/10.1109/TCYB.2018.2850750

119. F. Zhang, G. Shi, Co-representation network for generalized zero-shot learning, *International Conference on Machine Learning (ICML)*, (2019), 7434–7443.

120. H. Zhang, Y. Long, Y. Guan, L. Shao, Triple verification network for generalized zero-shot learning, *IEEE T. Image Process.*, **28** (2019), 506–517. https://doi.org/10.1109/TIP.2018.2869696

121. L. Zhang, T. Xiang, S. Gong, Learning a deep embedding model for zero-shot learning, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 2021–2030. https://doi.org/10.1109/CVPR.2017.321

122. Z. Zhang, V. Saligrama, Zero-shot learning via semantic similarity embedding, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2015), 4166–4174. https://doi.org/10.1109/ICCV.2015.474

123. H. Zheng, J. Fu, T. Mei, J. Luo, Learning multi-attention convolutional neural network for fine-grained image recognition, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2017), 5209–5217. https://doi.org/10.1109/ICCV.2017.557

124. Y. Zhu, J. Xie, B. Liu, A. Elgammal, Learning feature-to-feature translator by alternating back-propagation for generative zero-shot learning, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 9844–9854. https://doi.org/10.1109/ICCV.2019.00994

125. Y. Zhu, J. Xie, Z. Tang, X. Peng, A. Elgammal, Semantic-guided multi-attention localization for zero-shot learning, *Advances in Neural Information Processing Systems*, **32** (2019), 14917–14927.