Modality Independent Adversarial Network for Generalized Zero Shot Image Classification

Haofeng Zhang[a], Yinduo Wang[b,a], Yang Long[c], Longzhi Yang[d], Ling Shao[e]

[a] *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China*
[b] *The 29th Research Institute, China Electronics Technology Group Corporation, Chengdu, China*
[c] *School of Computer Science, Durham University, Durham, UK*
[d] *Department of Computer and Information Science, Northumbria University, Newcastle upon Tyne, UK.*
[e] *Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates*

Abstract

Zero Shot Learning (ZSL) aims to classify images of unseen target classes by transferring knowledge from source classes through semantic embeddings. The core of ZSL research is to embed both visual representation of object instance and semantic description of object class into a joint latent space and learn cross-modal (visual and semantic) latent representations. However, the learned representations by existing efforts often fail to fully capture the underlying cross-modal semantic consistency, and some of the representations are very similar and less discriminative. To circumvent these issues, in this paper, we propose a novel deep framework, called Modality Independent Adversarial Network (MIANet) for Generalized Zero Shot Learning (GZSL), which is an end-to-end deep architecture with three submodules. First, both visual feature and semantic description are embedded into a latent hyper-spherical space, where two orthogonal constraints are employed to ensure the learned latent representations discriminative. Second, a modality adversarial submodule is employed to make the latent representations independent of modalities to make the shared representations grab more cross-modal high-level semantic information during train-

Email addresses: zhanghf@njust.edu.cn (Haofeng Zhang), wangyd@njust.edu.cn (Yinduo Wang), yang.long@ieee.org (Yang Long), longzhi.yang@northumbria.ac.uk (Longzhi Yang), ling.shao@ieee.org (Ling Shao)

ing. Third, a cross reconstruction submodule is proposed to reconstruct latent representations into the counterparts instead of themselves to make them capture more modality irrelevant information. Comprehensive experiments on five widely used benchmark datasets are conducted on both GZSL and standard ZSL settings, and the results show the effectiveness of our proposed method.
Keywords: Generalized Zero Shot Learning (GZSL), Orthogonal Constraint, Cross Reconstruction, Adversarial Network, Modality Independent Learning

1. Introduction

Along with the development of deep learning techniques, conventional close set image classification has achieved the level of human beings. However, in this big data era, an increasing number of new categories of objects are emerging everyday, so conventional close set methods need to be retrained to include the new categories, which is time-consuming and infeasible for realistic applications. To circumvent this issue, Zero Shot Learning (ZSL) [54, 53, 1, 36, 58, 25] is proposed to recognize novel categories that are invisible during training. This task is achieved by the assist of an auxiliary intermediate information, *e.g.* attributes annotated by experts [13], to establish the bridge between the source and target objects, just as our human beings classify novel categories via the previous knowledge. In recent years, ZSL has achieved great success and attracted much attention, but traditional ZSL only concentrates on classifying novel objects within the scope of unseen categories, which is unreasonable in realistic scenarios because we cannot decide the ascription of the new emerging instance. Therefore, Chao *et al.* [7] proposed a more realistic Generalized ZSL (GZSL) setting, which extends the testing scope from only unseen classes to all classes, including both seen and unseen.

As shown in Fig. 1, most of the existing ZSL methods address this task by finding the relationships between visual features and semantic embeddings of seen classes and then transferring them to unseen categories. Therefore, how to make the model extract the most representative latent embeddings of seen

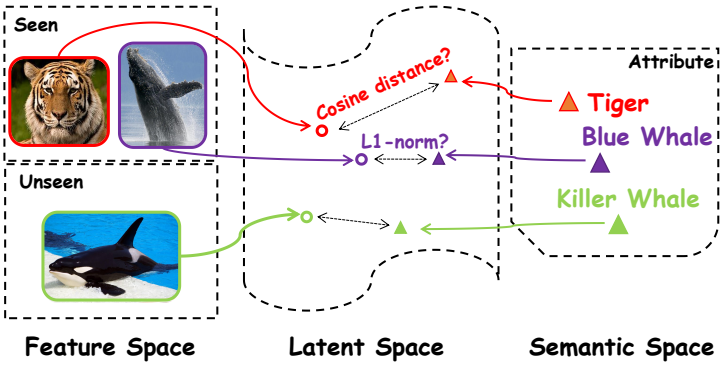


Figure 1: The core problem of ZSL is how to learn the most representative representations in latent space.

categories is one of the most crucial issues for ZSL research. Some early efforts such as Direct Attribute Projection (DAP) [25], Attribute Label Embedding (ALE) ALE [2] and Semantic Auto-Encoder (SAE) [23], have adopted different constraints, *e.g.*, cosine distance, bilinear compatibility function or Frobenius norm, to constrain the pairwise similarity between visual feature and semantic attribute during the training phase. However, visual features and semantic attributes are from two different modalities, where the modality discrepancy exists, thus the learned shared representation will fail to capture the underlying cross-modal semantic information, and finally lead to unsatisfactory classification results. We argue that employing a simple pairwise constraint is far from sufficient because it only considers the pairwise correlation, while the most essential information transferred to target domain should be the high-level semantic consistency, which is needed to be fully modeled.

Besides, attributes of some categories are very similar to each other, which make them less discriminative, *e.g* the attribute of 'Persian Cat' is similar to that of 'Siamese Cat', and consequently lead to a wrong classification. Zhang *et al.* in [52] tried to enlarge the gap between the attributes of seen classes by employing triple verifications. However, this method only concentrates on the prototypes of seen classes, while the unseen classes are totally ignored, which

makes some prototypes of unseen classes still similar to each other, and finally lead to a wrong classification, especially on the more realistic GZSL setting. In addition, some other efforts disperse the distance of class prototypes of all classes but only focus on semantic modality and ignore the correlation between visual features and semantic embeddings [21], which makes the latent representation less discriminative and less representative.

Recently, unseen sample synthetic based methods have attracted more attention due to their excellent performance [20]. Different from compatible methods [48], they often utilize Generative Adversarial Network (GAN) [16] to train a projection from semantic attributes to visual features, and then exploit the learned network model to synthesize samples of unseen classes, which are subsequently combined with seen samples to train a supervised close-set model [34]. However, these synthetic based methods usually encounter a serious problem, that is, they are learned within a close-set, and when a new class is added, the entire model should be retrained with new synthetic visual samples.

In order to solve the aforementioned problems, in this paper, we propose a novel deep framework, called Modality Independent Adversarial Network (MIANet) for generalized zero shot image classification, which is an end-to-end architecture to learn the more discriminative and representative latent representations. The novelties of our model lie in the following three aspects. Firstly, both visual features and semantic attributes are projected into a latent space, where two orthogonal constraints, including semantic to semantic and semantic to visual, are employed to make the latent representations more discriminative. Secondly, an adversarial training mechanism is constructed in the latent space. Considering the former projector as a generator to yield semantic discriminative latent representation, we employ another modality discriminator to distinguish the modalities of the shared representation, which competes with the generator to alternately boost each other. This adversarial training mechanism can facilitate the learned representation more discriminative for semantics but indistinguishable for modalities [16], thus it can effectively enhance the cross-modal semantic consistency and will finally lead to performance improvement. It is

noteworthy that this adversarial strategy is different from those GAN based unseen sample synthetic based methods [47, 31, 53], which often suffer from the retraining problem of the close-set image classification when new category emerges. At last, in order to alleviate the domain shift problem [15, 23] and let the latent vectors preserve more information from their respective original spaces, we proposed a reconstruction subnetwork. Instead of directly reconstructing themselves, we propose a novel cross reconstruction submodule to reconstruct the counterpart representation, which can preserve more cross-modal information. The contributions of this work are summarized as follows,

- We propose a novel and effective zero shot image classification framework, namely Modality Independent Adversarial Network (MIANet), which is an end-to-end architecture to learn more representative and discriminative latent representation for visual features and semantic embeddings;
- To create a discriminative latent space, two orthogonal constraints are applied to make each vector orthogonal to other if they belong to different categories, otherwise normalized. Furthermore, a cross reconstruction framework for both visual and semantic modalities is employed to make the latent vectors more representative;
- Adversarial training mechanism is exploited to confuse the source modality of the latent vectors, which can make the pairwise vectors indistinguishable for modalities, and results in preserving more high-level semantic consistency within them;
- Extensive experiments are conducted on five popular datasets, and the results show that our MIANet can outperform most of the state-of-the-art compatible methods on both GZSL and ZSL settings.

The main content of this paper is organized as follows: In section 2 we briefly introduce the existing methods for ZSL and GZSL. Section 3 describes the proposed method in detail. Section 4 gives the experimental results of

comparison with existing methods on several metrics. Finally in section 5, we conclude this paper.

2. Related work

2.1. Compatible methods

Zero Shot Learning (ZSL) models are trained without unseen classes but try to classify unseen samples within the scope of unseen categories. So far, many researchers have been devoting to this research area. Early efforts like DAP [25] estimate the labels by learning probabilistic attribute classifiers. In ALE [2] and SJE [3], Akata *et al.* projected visual features into semantic space via a bilinear compatibility constraint. Other ZSL baseline methods such as CONvex combination of Semantic Embeddings (CONSE) [35] and Semantic Similarity Embedding (SSE) [57] try to automatically build unseen attributes from the instances of seen categories to reduce the effect of manual attributes. Furthermore, researchers like Kodirov *et al.* [23] used the concept of Auto-Encoder and directly use Euclidean distance to constrain the similarity of projected visual vectors and semantic embeddings. After that Ding *et al.* in [10] made efforts on the method based on low-rank embedded semantic dictionary. These methods often employ a constraint to measure the similarity of latent vectors that from the same category. However, we argue that the high-level semantic consistency, which should be transferred to unseen classes, is the core information and employing such a manual constraint is insufficient. What's more, these methods only focus on the relationship between visual features and semantic embeddings from seen classes and do not take full advantage of the attributes of all the categories.

Different from conventional ZSL, which assumes that all the test samples are only from unseen categories, Generalized ZSL (GZSL), which is firstly proposed by Chao *et al.* in [7], enlarges the search scope to all classes, because we cannot obtain the information that whether the test data only belongs to the unseen classes beforehand in most scenarios, therefore GZSL is a more realistic and

challenging task. Besides, it is noteworthy that Xian *et al.* [48] in 2017 put forward a new split of several popular datasets for GZSL testing, and released a benchmark of some recent ZSL methods, which has greatly promoted the development of ZSL research. To compensate the domain gap between the seen classes and the unseen classes, many specialized methods have been developed for GZSL. For example, Zhang et al. proposed an embedding model called co-representation network [50], which contains two modules, one of them is a collaborative module for projecting the semantic space into the visual embedding space, and another is relationship module for classification. This model can learn a uniform visual embedding space that effectively alleviates the bias problem. Liu *et al.* proposed a novel Deep Calibration Network (DCN) [28] approach, which enables simultaneous calibration of deep networks on the confidence of source classes and uncertainty of target classes.

2.2. Synthetic based methods

Recently, synthetic based methods have elicited wide interest among researchers because they can achieve very significant performance. Long *et al.* in [31] firstly proposed to use the attributes of unseen classes to synthesize unseen visual features, and then train a fully supervised model with the seen data and the synthesized unseen visual features. From then on, an increasing number of synthesized feature based methods have being proposed [53, 24, 40, 8], and many of them are based on Variational Auto-Encoder (VAE) [22] or GAN [16] since adversarial learning can encourage the networks to synthesize more realistic samples [55, 56]. CVAE-ZSL [33] uses a conditional variational autoencoder to implement the unseen sample generation. Xian *et al.* proposed a method called f-CLSWGAN to train a Wasserstein GAN with a classification loss and is able to generate sufficiently discriminative CNN features [47]. Huang *et al.* [20] trained three components, including visual to semantic mapping, semantic to visual mapping and a metric to evaluate the closeness of an image feature and a class embedding, under the combination of cyclic consistency loss and dual adversarial loss to learn a visual generative network for unseen classes.

Dual Adversarial Semantics-Consistent Network (DASCN) [34] learns primal and dual GAN in a unified framework, where the primal GAN learns to synthesize inter-class discriminative and semantics-preserving visual features and the dual GAN enforces the synthetic visual features to represent prior semantic knowledge via semantics-consistent adversarial learning. Although these methods can achieve significant performance, they all suffer from a common serious problem that when there comes an object of a new category, the model should be retrained with the new synthesized samples of the coming category. Different from these GAN based synthetic methods, our approach is a compatible one, which does not suffer from the previous mentioned problem, and it utilizes adversarial training to generate latent modality independent vectors for both visual features and semantic attributes.

2.3. Representation Learning

Representation Learning is widely used in cross-modal retrieval methods. Since visual features and semantic attributes are from different modalities, and they usually have inconsistent representation and distribution, thus it is necessary to find a way to measure the semantic similarity of samples across modalities and learn the cross-modal representations to bridge the modality gap. A variety of cross-modal methods [12, 17, 38] have been proposed in different ways to learn the common representation in latent space. Early efforts [19] employs Canonical Correlation Analysis (CCA) to maximize pairwise correlation of cross-modal data to learn a linear projection matrices. Cross-modal Factor Analysis (CFA) [27] learns the the common representation of pairwise data by directly minimizing the Frobenius norm between them. Recently, Wang *et al.* in [44] for the first time utilize an adversarial training mechanism to minimize the gap among the representations from different modalities for image retrieval, which can effectively preserve the high-level semantic consistency in latent space.

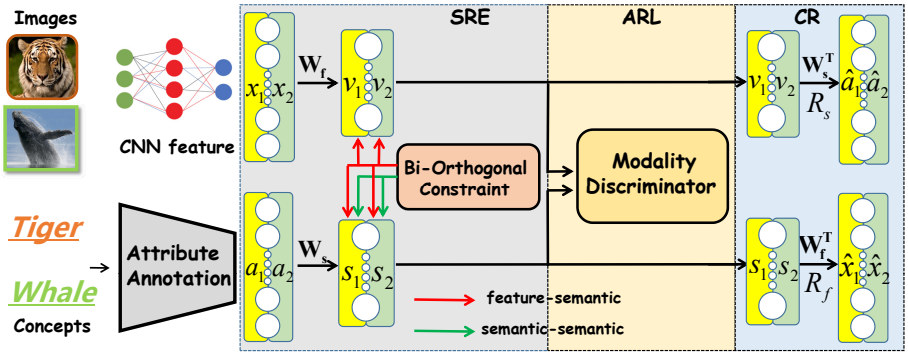


Figure 2: An illustration of the framework of our proposed method. "SRE" means shared representation embedding, "ARL" stands for adversarial representation learning and "CR" denotes cross reconstruction.

3. Methodology

3.1. Problem Definition

Given a dataset $\mathcal{D}$, which is composed of two disjoint groups, seen classes $\mathcal{S} = \{1, \cdots, s\}$ and unseen classes $\mathcal{U} = \{s+1, \cdots, s+u\}$, where $\mathcal{S} \cap \mathcal{U} = \varnothing$. The training set $\boldsymbol{X}_s \in \mathbb{R}^{d \times N}$ is consist of N d-dimensional visual features and each of them is associated with a label from seen classes, while K d-dimensional visual features of unseen classes construct the testing set $\boldsymbol{X}_u \in \mathbb{R}^{d \times K}$, which is invisible during training phase. As for the auxiliary intermediate semantic embeddings, it is given a set $\boldsymbol{A}_{all} = \boldsymbol{A}_s \cup \boldsymbol{A}_u$, where $\boldsymbol{A}_s \in \mathbb{R}^{l \times s}$ denotes the corresponding class-level s l-dimensional attributes of seen classes, while $\boldsymbol{A}_u \in \mathbb{R}^{l \times u}$ represents the unseen categories attributes. Under both the conventional ZSL setting and the more realistic GZSL setting, $\boldsymbol{X}_s$, $\boldsymbol{A}_s$ and $\boldsymbol{A}_u$ are assumed to be known in advance, but the difference is that the goal of ZSL is to recognize unseen samples $\boldsymbol{X}_u$ with the search scope fixed on $\mathcal{U}$, while GZSL executes classification on both $\mathcal{U}$ and $\mathcal{S}$.

3.2. Model Architecture

The proposed MIANet aims at learning more discriminative and representative cross-modal latent representations of visual features and semantic attributes, and the whole architecture is illustrated in Fig. 2. MIANet consists of three subnetworks: 1) Shared Representation Embedding (SRE) subnetwork (gray block in Fig. 2), 2) Modal-Adversarial Representation Learning (ARL) subnetwork (yellow block) and 3) Cross Reconstruction (CR) subnetwork (blue block). The three subnetworks form an end-to-end deep architecture, which is alternately trained to generate cross-modal latent representations of both visual features and semantic attributes.

3.2.1. Shared Representation Embedding

The Shared Representation Embedding (SRE) subnetwork is proposed to perform projection from visual space and semantic space into latent space. During the training stage, we take one visual feature $\boldsymbol{x}_i$ and one semantic attribute $\boldsymbol{a}_i$ as a pair $< \boldsymbol{x}_i, \boldsymbol{a}_i >$, as shown in the upper-left and lower-left parts in Fig. 2, then feed them simultaneously into two independent fully connected networks with the parameters $\boldsymbol{W}_f$ and $\boldsymbol{W}_s$ respectively to learn the latent representations.

Due to the fact that some manually annotated attributes and visual features of different categories are very similar to each other, *e.g*, the attribute of 'Persian Cat' is very similar to that of 'Siamese Cat'. Thus it is necessary to make the representations far away from each other if they come from different categories while ensuring that the representations of the same category are closer in latent space. Here, we constrain the projected representations should be normalized in latent space, so the vectors in latent space can be considered as the distribution points on the surface of a unit hyper-sphere. In addition, it is impossible to make the representations of all categories far away from each other on the limited surface of the sphere, thus the best way to make them discriminative is to let them orthogonal with each other. Here, we propose a Bi-Orthogonal Constraint (BOC) in this latent space to fulfill the above demands.

BOC consists of two different orthogonal constraints, one is for feature-semantic, shown as the red arrows in Fig. 2, and another is for semantic-semantic, shown as the green arrows in Fig. 2. The former one builds up a bridge between the visual and semantic modalities, and the latter one is to take full advantages of the attributes of both the seen and unseen classes to make the latent semantic representations more discriminative for all classes, which is very beneficial for improving the classification performance, especially for the more realistic GZSL setting.

Firstly, for the feature-semantic orthogonal constraint, we randomly select a visual-semantic pair $< \boldsymbol{x}_i, \boldsymbol{a}_i >$ and feed it into the two independent networks shown in Fig. 2, then two latent representations $\boldsymbol{v}_i$ and $\boldsymbol{s}_i$ for visual feature $\boldsymbol{x}_i$ and the semantic attribute respectively can be obtained. Here, when $\boldsymbol{x}_i$ and $\boldsymbol{a}_i$ come from different classes, the similarity score (cosine distance) of them is set to zero, that is to say, constraining them to be orthogonal to each other, and contrarily, it is set to one when they belong to the same category. Therefore, the loss function of this feature-semantic orthogonal constraint can be represented with Cross-Entropy,

$$\mathcal{L}_{orf} = -\frac{1}{n}\sum_{i=1}^{n}(m_i log(\boldsymbol{v}_i^T \boldsymbol{s}_i) + (1 - m_i) log(1 - \boldsymbol{v}_i^T \boldsymbol{s}_i)), \tag{1}$$

where, n is the batch size used in training, m_i is the ground-truth similarity label of each pair $< \boldsymbol{x}_i, \boldsymbol{a}_i >$, when $\boldsymbol{x}_i$ and $\boldsymbol{a}_i$ share the same class label, m_i equals 1, otherwise it equals 0.

Another orthogonal constraint is employed for semantic-semantic discrimination. It is hoped that all the latent representations of semantic attributes, including both the seen and the unseen, are discriminative, thus we feed all the class attributes $\boldsymbol{A}_{all}$ into the semantic subnetwork to obtain their latent representations $\boldsymbol{A}_l$. The best way to make them discriminative is to constrain them to be orthogonal to each other on the limited surface of the unit hyper-sphere. we directly constrain their inner product instead of the cosine similarity, this constraint will guide them to be orthogonal for each other while also make these vectors to be normalized. The loss function of the semantic-semantic orthogonal

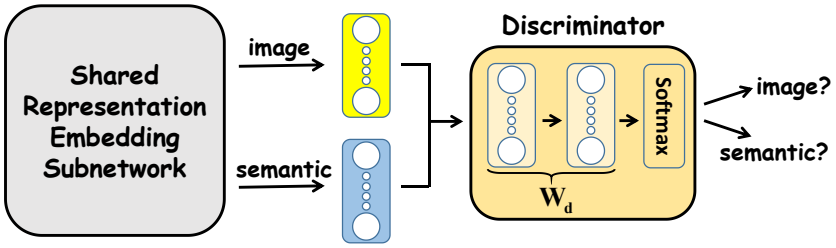


Figure 3: The detail structure of our Modality-Adversarial Representation Learning Subnetwork.

constraint can be represented as,

$$\mathcal{L}_{ors} = \|\boldsymbol{A}_l^T \boldsymbol{A}_l - \boldsymbol{I}\|_F^2. \tag{2}$$

3.2.2. Modal-Adversarial Representation Learning

This subnetwork is proposed to further discover the cross modality relationships between two latent representations. Although the introduced SRE subnetwork has already learned latent representation of visual features and semantic attributes, there are still two problems,

1) There are two independent subnetworks for visual and semantic modalities, and they only consider whether they come from the same class or not, which cannot fully extract the modality-irrelevant information to reduce the heterogeneity gap.

2) Like many existing ZSL efforts, SRE uses a predefined metric to extract the pairwise relationship, but ignores the high-level semantic consistency, which should be effectively contained.

It is believed that ideal latent representations should have two properties, and both of them are very important for the information transferred between domains and modalities in ZSL task. One is that the latent representation should be discriminative for semantics, so it can preserve more semantic consistency within a single modality. Another is that the representation should also

be indistinguishable for modalities, so the heterogeneity gap can be effectively reduced. Therefore, for addressing the above problems and inspired by [44], in this paper we build an adversarial training mechanism to confuse whether the latent representation comes from the visual space or the semantic space, and train our model to reduce the modality discrepancy to make the representations discriminative for semantics but indistinguishable for modalities.

The architecture of this subnetwork is shown in Fig. 3. It considers the former subnetwork as a latent representation generator to yield semantic discriminative vectors. Then we build another modality discriminator to distinguish different modalities, that is to say, the aforementioned generator aims to maximize the semantic discriminative ability, while the discriminator is proposed to maximize the modality differentiable ability. To be specific, the discriminator is consisted of two fully connected layers with the parameters $\boldsymbol{W}_d$, and followed by a softmax layer as the output. We feed both the latent representations of $\boldsymbol{v}_i$ and $\boldsymbol{s}_i$ into it, where each input vector is assigned with a binary ground-truth value to indicate which modality it belongs to. The loss function of this discriminator can be represented with Cross-Entropy,

$$\mathcal{L}_{dis} = \frac{1}{2n}(\sum_{i=1}^{n} f(\boldsymbol{y}_i^{(x)}, D(\boldsymbol{v}_i)) + \sum_{i=1}^{n} f(\boldsymbol{y}_i^{(s)}, D(\boldsymbol{s}_i)), \tag{3}$$

where, $D(\cdot)$ denotes the discriminator and it outputs 2-dimensional probability value for modality indicator. $\boldsymbol{y}_i^{(x)}$ and $\boldsymbol{y}_i^{(s)}$ are the corresponding modality ground-truth for $\boldsymbol{x}_i$ and $\boldsymbol{a}_i$, and they are represented in the form of one-hot vector, *i.e.*, the ground-truth label of $\boldsymbol{y}_i^{(x)}$ is 10, and that of $\boldsymbol{y}_i^{(s)}$ is 01. $f(\cdot, \cdot)$ is the Cross-Entropy function, which is defined as,

$$f(\boldsymbol{y}, \boldsymbol{d}) = -\sum_{i=1}^{2} y_i log d_i + (1 - y_i) log(1 - d_i), \tag{4}$$

where, y_i and d_i are the ith entry of $\boldsymbol{y}$ and $\boldsymbol{d}$. During training stage, the modality discriminator aims to distinguish different modalities, while the latent representations generator tries to decrease the discrepancy of cross-modal representation to confuse the modality discriminator through an adversarial training strategy.

3.2.3. Cross Reconstruction

At last, in order to reduce the influence of the projection domain shift problem [15] and make the latent vectors more representative, we proposed a Cross Reconstruction (CR) subnetwork to reconstruct the original vectors $\boldsymbol{x}_i$ and $\boldsymbol{a}_i$ from the latent representation $\boldsymbol{v}_i$ and $\boldsymbol{s}_i$, which means that the latent representations can carry more information from their respective original spaces. Due to the function of discriminator learning, the latent representation $\boldsymbol{v}_i$ and $\boldsymbol{s}_i$ so far contain the information of not only their original modality but also the other modalities. Therefore, different from traditional AutoEncoder, we propose CR to hope that the latent vectors can preserve more cross-modal information in addition to the original information.

To be specific, we constrain the latent vectors can be reconstructed to the counterpart representation, *i.e* the original semantic vector should be able to be rebuilt by the latent visual feature representation and vice versa. The reconstruction loss can be defined as, with the Frobenius norm,

$$\mathcal{L}_{rec} = \|R_s(\boldsymbol{v}_i) - \boldsymbol{a}_i\|_F^2 + \|R_f(\boldsymbol{s}_i) - \boldsymbol{x}_i\|_F^2, \tag{5}$$

where, the function $R_f(\cdot)$ and $R_s(\cdot)$ denote the feature reconstructor and semantic reconstructor respectively, which can be found in Fig. 2, and the parameters of this subnetwork are the transpose of corresponding embedding parameters $\boldsymbol{W}_f$ and $\boldsymbol{W}_s$.

Our MIANet is an end-to-end architecture with three submodules to learn more discriminative and representative latent vectors for both visual features and semantic attributes, and the three subnetworks are trained in two adversarial processes to boost each other. In the testing phase, *e.g.* under GZSL setting, each test image feature and attributes of all classes are fed into the network to generate the latent representations. Then we directly calculate the similarities, such as the cosine distance, of them to find the nearest pair as their classification result.

Table 1: Summary of the five employed datasets. "SS" denotes the number of Seen Samples for training, "TS" and "TR" refer to the numbers of unseen class samples and seen class samples respectively for testing.

Datasets	Dimension		Class Number		Samples Number		
	Feat.	*Att.*	*Seen*	*Unseen*	*SS*	*TS*	*TR*
SUN	2048	102	645	72	10320	1440	2580
CUB	2048	312	150	50	7057	2967	1764
AWA1	2048	85	40	10	19832	4958	5685
AWA2	2048	85	40	10	23527	5882	7913
aPY	2048	64	20	12	5932	7924	1483

3.3. Optimization

Since the proposed model is designed to generate modality-irrelevant representations for both visual features and semantic attributes, it should be trained through an adversarial manner like a two-player game with an iterative manner,

1) Fix $\boldsymbol{W}_f$ and $\boldsymbol{W}_s$, and update $\boldsymbol{W}_d$ only, which can be written as,

$$\widehat{\boldsymbol{W}}_d = arg \min_{\boldsymbol{W}_d} \mathcal{L}_{dis}. \tag{6}$$

After this operation, the discriminator can distinguish the latent representations of different modalities better.

2) By fixing the discriminator parameters $\boldsymbol{W}_d$, the generators and the reconstructors can be updated by,

$$(\widehat{\boldsymbol{W}}_f, \widehat{\boldsymbol{W}}_s) = arg \min_{\boldsymbol{W}_f, \boldsymbol{W}_s} \mathcal{L}_{orf} + \alpha \mathcal{L}_{ors} + \beta \mathcal{L}_{rec} - \gamma \mathcal{L}_{dis}, \tag{7}$$

where, maximizing $\mathcal{L}_{dis}$ can encourage our model to reduce the heterogeneity gap among modalities and generate more similar representations to confuse the discriminator, so that our model can grab more high-level semantic consistence information. In addition, minimizing other losses can make the latent representation more discriminative.

4. Experiments

4.1. Datasets

In our experiments, we employ five popular benchmark datasets, *i.e.*, SUN (SUN attribute) [37], CUB (Caltech-UCSD-Birds 200-2011) [43], AWA1 (Animals with Attributes) [26], AWA2 and aPY (attribute Pascal and Yahoo) [11]. Among them, SUN and CUB are fine-grained while AWA1/2 and aPY are coarse-grained. Some other details of the datasets can be found in Tab. 1, where "SS" denotes the number of Seen Samples for training, "TS" and "TR" refer to the numbers of unseen class samples and seen class samples respectively for testing. In addition, all the comparisons with state-of-the-art methods below are under the same split strategy, which is proposed by Xian *et al.* in [48].

Table 2: Comparison of our MIANet and state-of-the-art methods under GZSL setting. Bold font stands for the best result of the corresponding column and '-' means not reported.

	SUN			**CUB**			**AWA1**			**AWA2**			**APY**		
Method	*ts*	*tr*	*H*	*ts*	*tr*	*H*	*ts*	*tr*	*H*	*ts*	*tr*	*H*	*ts*	*tr*	*H*
DAP [25]	4.2	25.1	7.5	1.7	67.9	3.3	0.0	**88.7**	0.0	0.0	84.7	0.0	4.8	78.3	9.0
CONSE [35]	6.8	**39.9**	11.6	1.6	**72.2**	3.1	0.4	88.6	0.8	0.5	**90.6**	1.0	0.0	**91.2**	0.0
CMT [41]	8.1	21.8	11.8	7.2	49.8	12.6	0.9	87.6	1.8	0.5	90.0	1.0	0.0	91.2	0.0
LATEM [46]	14.7	28.8	19.5	15.2	57.3	24.0	7.3	71.7	13.3	11.5	77.3	20.0	0.1	73.0	0.2
SSE [57]	2.1	36.4	4.0	8.5	46.9	14.4	7.0	80.5	12.9	8.1	82.5	14.8	0.2	78.9	0.4
ALE [2]	21.8	33.1	26.3	23.7	62.8	34.4	16.8	76.1	27.5	14.0	81.8	23.9	4.6	73.7	8.7
DEVISE [14]	16.9	27.4	20.9	23.8	53.0	32.8	13.4	68.7	22.4	17.1	74.7	27.8	4.9	76.9	9.2
SJE [3]	14.7	30.5	19.8	23.5	59.2	33.6	11.3	74.6	19.6	8.0	73.9	14.4	3.7	55.7	6.9
ESZSL [39]	11.0	27.9	15.8	12.6	63.8	21.0	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6
SYNC [6]	7.0	43.4	13.4	11.5	70.9	19.8	8.9	87.3	16.2	10.0	90.5	18.0	7.4	66.3	13.3
SAE [23]	8.8	18.0	11.8	7.8	54.0	13.6	1.8	77.1	3.5	1.1	82.2	2.2	1.1	82.2	2.2
GFZSL [42]	0.0	39.6	0.0	0.0	45.7	0.0	1.8	80.3	3.5	2.5	80.1	4.8	0.0	83.3	0.0
LAGO [4]	18.8	33.1	23.9	21.8	73.6	33.7	23.8	67.0	35.1	-	-	-	-	-	-
PSEUDO [30]	19.0	32.7	24.0	23.0	51.6	31.8	22.4	80.6	35.1	-	-	-	15.4	71.3	25.4
KERNEL [51]	21.0	31.0	25.1	24.2	63.9	35.1	18.3	79.3	29.8	18.9	82.7	30.8	11.9	76.3	20.5
TRIPLE [52]	18.2	28.9	22.3	26.5	62.3	37.2	27.0	67.9	38.6	28.5	66.7	39.9	16.1	66.9	25.9
VZSL [45]	15.2	23.8	18.6	17.1	37.1	23.8	22.3	77.5	34.6	21.7	78.6	34.0	8.4	75.5	15.1
LESAE [29]	21.9	34.7	26.9	24.3	53.0	33.3	19.1	70.2	30.0	21.8	70.6	33.3	12.7	56.1	20.1
LESD [9]	15.2	19.8	17.2	14.6	38.5	21.2	12.6	71.0	21.4	15.3	71.5	25.2	11.8	49.3	19.0
Ours	**22.2**	35.6	**27.4**	**33.3**	49.5	**39.9**	**46.5**	68.5	**55.4**	**43.7**	70.2	**53.3**	**27.6**	55.8	**37.0**

4.2. Experimental Setting

We employ the extracted features with ResNet [18] as our input, and all the settings, *e.g.* employed attributes and classes split, are the same as that in [48]. Additionally, since the probability of randomly selecting one pair of visual

feature and semantic attribute that belong to the same category is extremely low, we employ a strategy to build training pairs to ensure that the positive and negative pairs are balanced in quantity. Concretely, we first attach an attribute of the same category to each training feature as a positive pair, and then we construct a negative pair by attaching an attribute of a random different class to the feature.

As for the details of the SRE architecture, we deploy a three-layer Fully-Connected (FC) deep network with ReLU activation to nonlinearly project the raw image features into the latent space and a two-layer FC network to embed attributes, and the parameter numbers of the networks are $2048 \rightarrow 2048 \rightarrow 2 \times (s+u)$ for visual features and $l \rightarrow 2 \times (s+u)$ for semantic attributes respectively. The CR subnetwork employs the transpose layers and parameters of the SRE. Besides, the discriminator is also composed of a two-layer FC network, and the layer dimensions are $100 \rightarrow 30 \rightarrow 2$ with ReLU activation attached to the middle layer.

There are three hyper-parameters α, β, γ in our method, and we randomly select 20% of the seen classes in 'SS' as validation unseen classes, and the parameters of the best average performance of 5 executions are picked as the optimal parameters. The optimal values of α is obtained as 100 on SUN and 1000 on other datasets, and β, γ are obtained as 0.01, 0.01 respectively. The great difference of which is mainly due to that the entropy loss value and Frobenius norm loss values have different orders of magnitude. At last, we set the batch size to 300 and the learning rate to 1×10^{-4}. In addition to some of the baseline methods evaluated in [48], we also compare our MIANet with some newly proposed frameworks, such as GFZSL [42], LAGO [4], PSEUDO [30], KERNEL [51], TRIPLE [52], LESAE [29], LESD [9] and VZSL [45].

4.3. Results on GZSL

Since GZSL is a more realistic and valuable setting than conventional ZSL, we firstly discuss the performance on GZSL. The evaluation criteria employed for evaluating our model under GZSL setting is the harmonic mean H, which

Table 3: Comparison of our MIANet and state-of-the-art methods under ZSL setting. Bold font stands for the best result of the corresponding column, and '-' means not reported.

Method(%)	**SUN**	**CUB**	**AWA1**	**AWA2**	**aPY**	**Average**
DAP [25]	39.9	40.0	44.1	46.1	33.8	40.8
CONSE [35]	38.8	34.3	45.6	35.9	26.9	36.3
CMT [41]	39.9	24.6	39.5	37.9	28.0	34.0
LATEM [46]	55.3	49.3	55.1	55.8	35.2	50.1
SSE [57]	51.5	43.9	60.1	61.0	34.0	50.1
ALE [2]	58,1	54.9	59.9	62.5	39.7	55.0
DEVISE [14]	56.5	52.0	54.2	59.7	39.8	52.4
SJE [3]	53.7	53.9	65.6	61.9	32.9	53.6
ESZSL [39]	54.5	53.9	58.2	58.6	38.3	52.7
SYNC [6]	56.3	55.6	54.0	46.6	23.9	47.3
SAE [23]	40.3	33.3	43.0	54.1	8.3	35.8
GFZSL [42]	**62.5**	42.0	55.6	63.8	32.8	51.3
LAGO [4]	57.5	57.8	-	64.8	-	-
PSEUDO [30]	60.4	57.2	66.2	-	40.4	-
TRIPLE [52]	59.3	54.9	64.7	65.8	40.9	57.1
VZSL [45]	52.0	43.8	63.7	64.2	30.3	50.8
LESD [9]	50.4	38.9	53.4	55.8	29.8	43.7
LESAE [29]	60.0	53.9	66.1	68.4	40.8	57.8
Ours	60.5	**57.9**	**70.1**	**69.0**	**41.2**	**59.5**

is defined as,

$$H = \frac{2 \times acc_{tr} \times acc_{ts}}{acc_{tr} + acc_{ts}}, \tag{8}$$

where, acc_{tr} and acc_{ts} are the accuracies of test samples from seen classes and unseen categories respectively, and we adopt the average per-class top-1 accuracy as the final result.

The experimental results on all five datasets are recorded in Tab. 2. From Tab. 2, we can clearly see that our MIANet outperforms all the other methods

on ts and H, which are two important metrics in GZSL setting [48]. To be specific, our MIANet improves H by 1.1% on SUN, 2.7% on CUB, 14.8% on AWA1, 22.5% on AWA2 and 11.1% on APY respectively. Compared to those existing methods that have high tr but low ts and H, such as DAP and CONSE, our MIANet can obtain more balanced results on ts and tr and eventually get a significant improvement on H. We ascribe this improvement to the Orthogonal Constraint and the Modal-Adversarial submodules, the former makes the latent space more discriminative for all the classes and the latter encourages the latent representations to preserve more high-level cross-modal information.

4.4. Results on ZSL

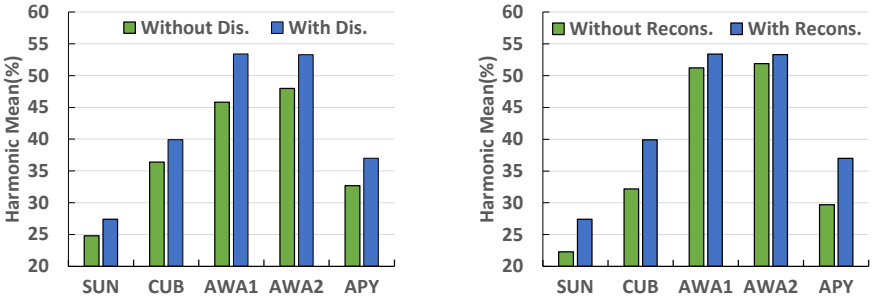
To further show the priority of our method, we also evaluate our MIANet under the conventional ZSL setting, where the search space is restricted on unseen classes. Similar with the experiment on GZSL, we adopt the average per-class top-1 accuracy as the final accuracy for ZSL, and the final results are recorded in Tab. 3. It is obvious that our MIANet can outperform other state-of-the-art methods on CUB, AWA1, AWA2 and APY, and also obtain competitive results on SUN. Specifically, we obtain the improvement of 0.1% on CUB, 3.9% on AWA1, 4.2% on AWA2, 0.3% on APY. Although our method do not perform the best in every dataset, we believe that a good method should perform well on most datasets rather than just on a single dataset, thus we also compare the average performance on all the five datasets, and the results are recorded in the last column of Tab. 3. It is clear that our MIANet can obtain the best average performance, which indicates the robustness of our method.

4.5. Detailed Analysis

4.5.1. Effect of Each Submodule

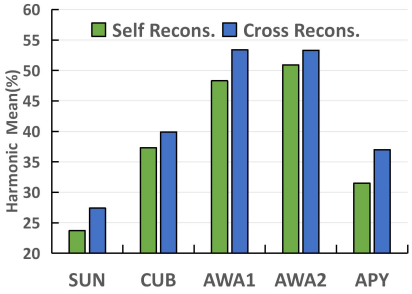
In this subsection, we conduct experiments to show how much the adversarial training mechanism and the cross reconstruction affect the final performance.

Firstly, we remove the ARL to investigate the effect of this submodule, and the results under GZSL setting are illustrated in Fig. 4 (a). It is clearly seen



(a) With or Without Adversarial

(b) With or Without Reconstruction

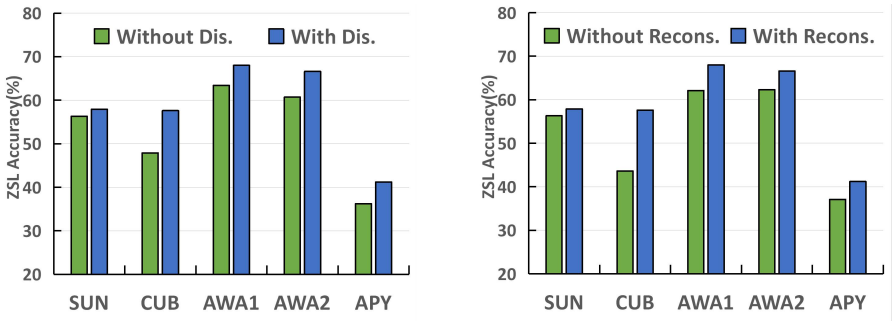


(c) Self or Cross Reconstruction

Figure 4: The effect of each submodule on the GZSL setting.

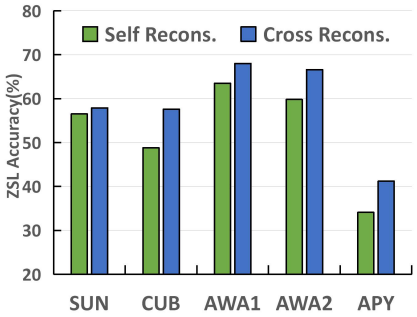
that the model with adversarial training mechanism can significantly enhance the final classification accuracy on all five datasets, especially on AWA1 and AWA2. In addition, the results under ZSL setting are illustrated in Fig. 5 (a), from which we can also find that this strategy can enhance the ZSL accuracy on all datasets. Since this adversarial training is employed to confuse the source modality of visual and semantic inputs, the projected visual vector can retain more global semantic information like attributes rather than the local details. Therefore, we attribute this improvement to the adversarial training mechanism for its ability of capturing high-level semantic consistency and modality independent representation.

Secondly, in order to analyze the effect of the CR subnetwork, we train our model with and without this submodule and conduct the test under GZSL set-



(a) With or Without Adversarial

(b) With or Without Reconstruction



(c) Self or Cross Reconstruction

Figure 5: The effect of each submodule on the ZSL setting.

ting, the results are shown in Fig. 4 (b). From this figure, it can be discovered that this constraint has a great impact on SUN, CUB and APY, because the reconstruction constraint can encourage the latent representations to preserve more information from their original space and also can alleviate the domain shift problem [23]. And the ZSL results are illustrated in Fig. 5 (b). The performance with reconstruction is better than that without it, especially on CUB, because CUB is a fine-grained dataset, the semantic information of which is more similar and needs more cross-modal information learned from CR subnetwork.

Thirdly, we replace the CR subnetwork with a self reconstruction module to show whether the cross reconstruction is effective. The experimental results on GZSL setting are recorded in Fig. 4 (c). Combined with Fig. 4 (b), we can see

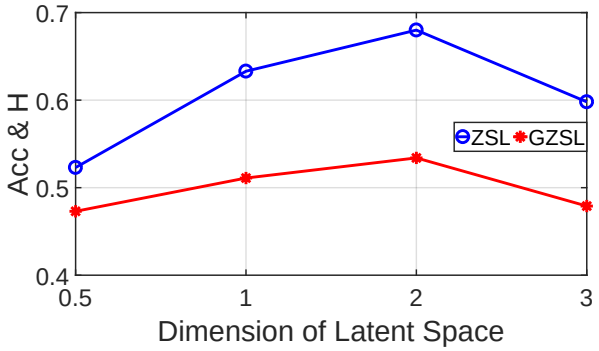


Figure 6: Comparison with different dimensions of the latent space on AWA1.

that although self reconstruction is better than no reconstruction, it is still not as good as the proposed CR, because reconstructing the cross-modal counterparts stimulates the latent representation to obtain more modality independent information. This phenomenon can be also found on other datasets like SUN, CUB, and AWA1 under ZSL setting in Fig. 5 (c).

4.5.2. Dimensions of Latent Space

In this subsection, we conduct experiment to show whether the dimension of the latent hyper-spherical space has an effect on the final classification accuracy. The results of both ZSL and GZSL on AWA1 are illustrated in Fig. 6, where the X-axis represents the multiple of the dimension of the latent space relative to the number of categories, and the blue line denotes the result on convention ZSL setting while the red one represents that on the more realistic GZSL setting. From Fig. 6, it is obvious that there is a common characteristic on both ZSL and GZSL, *i.e.*, when the number of dimensions of the latent space is about twice the number of categories, the ZSL accuracy and H can reach the peak. Since it is known that we need at least the same dimension as the category number to make all classes orthogonal to each other in latent space, the dimension cannot be too low. Besides, too high dimension may bring too much redundant information.

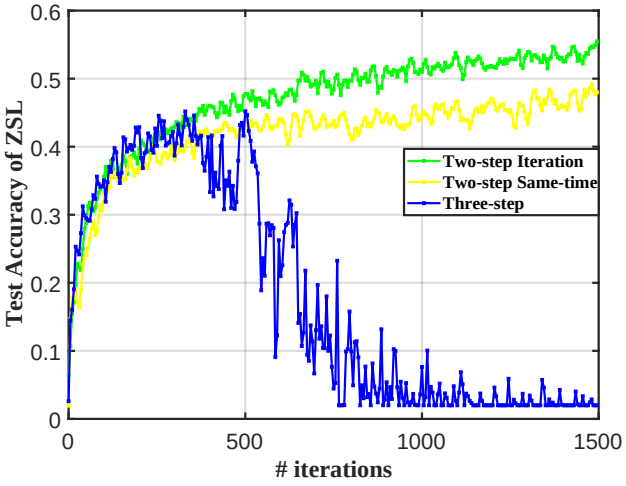


Figure 7: Test accuracy of ZSL with different strategy of optimization.

4.5.3. Effectiveness of Iterative Optimization Strategy

In this subsection, we conduct experiment to show the effectiveness of optimization via an iterative manner. To be specific, we compare the trend of ZSL accuracy under three different training strategy: 1) Green curve in Fig. 7, training with two optimizers, one is for the weights of discriminator by Eq. 6 and one is for the others by Eq. 7, and update the weights in an iterative manner. 2) Yellow curve of Fig. 7, two optimizers and we update the weights at the same time. 3) Blue curve in Fig. 7, we add a third optimizer to update all weights by Eq. 7. From Fig. 7, it can be clearly found that training with our iterative strategy can converge faster than that of updating weights at the same time, and obtain better results; the optimization will become easier when the discriminator or generator is well-trained. And if we update all weights at the same time by Eq. 7, the model will be confused and not sure which direction to make the optimization, and eventually lead to a bad result.

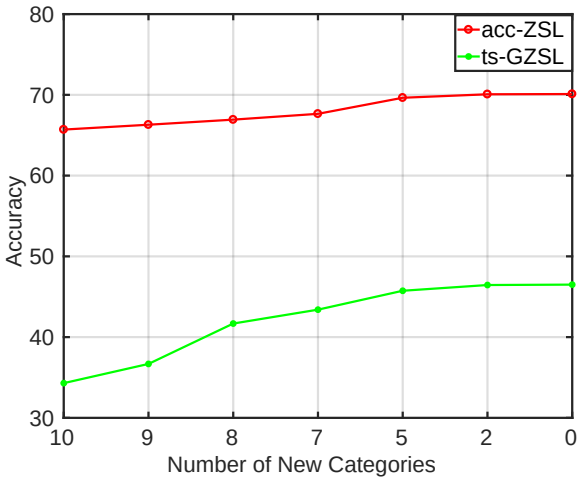


Figure 8: Test accuracy with respect to the number of new categories.

4.6. Emergence of new categories

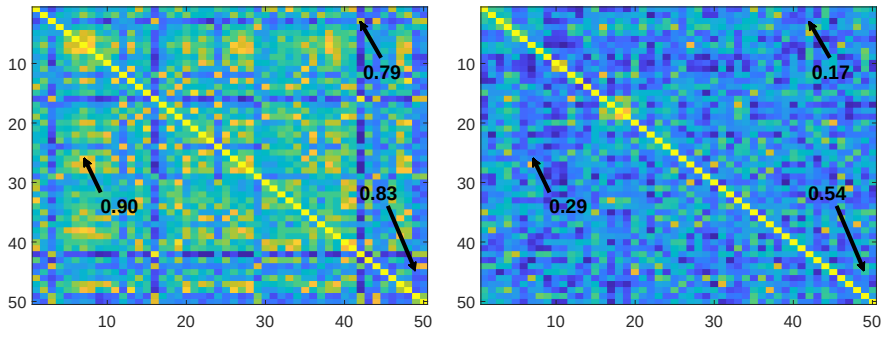
Since we have claimed that our method is different from those synthetic based method and can achieve competing performance when a new category appears, we have to conduct experiment to show how it performs when a new category emerges. In this experiment, we exploit the dataset AWA1 as an example, gradually decrease the number of unseen semantic attributes involved in training, and test the accuracies of the uninvolved categories on both ZSL and GZSL. We record the experimental results in Fig. 8, from which it can be clearly discovered that our method can still accept new categories and achieve competing performance, which is different from the synthetic based method that cannot accept any new category without retraining. The accuracy curves illustrated in Fig. 8 monotonically increase when the number of new categories decreases, which means that the more semantic attributes are involved in training the better performance our method can obtain.

4.7. Similarities of Class Prototypes

Since it is known that the more distinguishable the class prototypes are from each other, the easier the data samples can be classified, we analyze the similarities of class prototypes in attribute space and latent space respectively in this subsection. Firstly, we compute the normalized cosine similarity of each class prototype in both attribute and latent spaces on AWA1, and visualize the similarity matrix in Fig. 9. Specifically, the vectors from 0# to 40# in the matrix are the seen classes prototypes and the remaining ones belong to the unseen classes. Fig. 9 (a) demonstrates the original expert-annotated attributes similarity in semantic space, while Fig. 9 (b) illustrates the class prototype similarities learned with our MIANet in latent space. By comparing these two figures, it can be obviously discovered that the prototypes we learned are much more discriminative from each other, which reveals the effectiveness of our proposed method. Noted that not only seen classes become more discriminative against seen, but also seen against unseen and unseen against unseen become more discriminative. For example, we choose three pairs from the prototypes in Fig. 9, and the left one is the similarity of *'weasel'*(Seen) and *'hamster'*(Seen), the top-right one denotes the similarity of *'Killer Whale'* (Seen) and *'Blue Whale'* (Unseen) and the bottom-right one stands for *'Walrus'* (Unseen) to *'Seal'* (Unseen). Each of the three pairs is very similar to each other, and the similarities are all over 0.78 in cosine similarity, and it is difficult to classify them directly in original attribute space, while our model can make them much more different and the similarities of them are substantially decreased, which shows the superiority of our proposed method.

4.7.1. Distribution in Latent space

The objective of the proposed orthogonal constraint in latent space is to disperse all the classes, including both seen and unseen, and make them more discriminative. Therefore, in order to have a more intuitive understanding, we employ t-SNE [32] to illustrate the distributions of AWA1 in this space. Specifically, we choose several representative class pairs whose cosine similarities



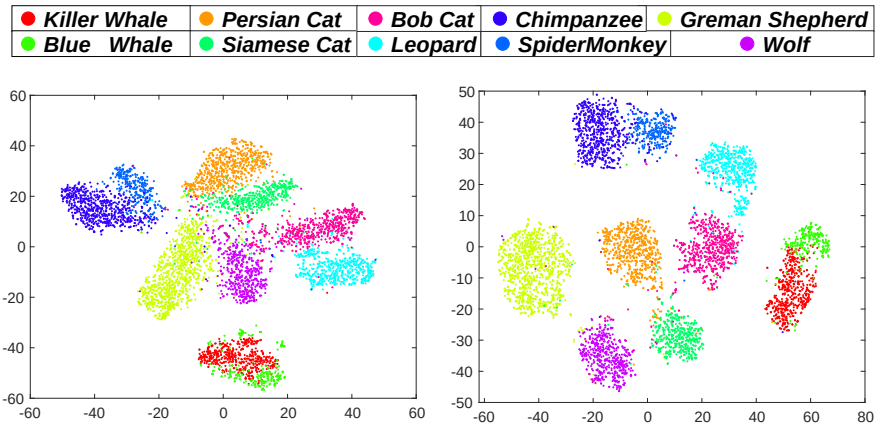
(a) Semantic space (b) Latent space

Figure 9: The cosine similarities of class prototypes in latent space on AWA1. Best viewed in color.

of prototypes in semantic space is very high, *i.e.*, they are very similar and hard to be classified. After the process, we finally obtain five pairs, including eight seen classes and two unseen classes, which can be found in the legend of Fig. 10. In Fig. 10, we illustrate the data distributions of the samples from the selected classes with and without the orthogonal constraint in the latent space. From this figure, it can be clearly seen that the samples of 'Killer Whale' (Seen) and 'Blue Whale' (Unseen) are overlapped without orthogonal constraint, while our MIANet can disperse them effectively. This phenomenon can also be found in other pairs, such as 'Persian Cat' and 'Siamese Cat', which indicates our method can perform well for all the classes.

4.7.2. Zero Shot Image Retrieval

In this subsection, we conduct experiments to show zero shot retrieval performance of our proposed MIANet. In this task, we apply the semantic attributes of each unseen category as the query vector, and compute the mean Average Precision (mAP) of the returned images. MAP is a popular metric for evaluating the retrieval performance, it comprehensively evaluates the accuracy and



(a) without Orthogonal Constraint (b) with Orthogonal Constraint

Figure 10: Visualization of similar classes on AWA1 in latent space with t-SNE [32]. Best viewed in color.

ranking of returned results, and defined as,

$$mAP = \frac{1}{u}\sum_{i=1}^{u}\left(\frac{1}{r_i}\sum_{j=1}^{r_i}\frac{j}{p_i(j)}\right), \tag{9}$$

where, r_i is the number of returned correct images from the dataset corresponding to the ith query attribute, $p_i(j)$ represents the position of the jth retrieved correct image among all the returned images according to the ith query attribute. In this experiment, the number of returned images equals the number of the samples in unseen classes.

For the convenience of comparison, we employ the standard split of the four datasets, including SUN, CUB, AWA1 and aPY, which can be found in [48], and the results are shown in Tab. 4. The values of the baseline methods listed in Tab. 4 are directly cited from [10]. The results show that our method can outperform the baselines on all four datasets, especially on the fine-grained dataset CUB, which reveals that our method can make the prototypes in latent space more discriminative.

Furthermore, we randomly select five unseen class attributes from AWA2 [48] as the query vectors, and the returned top-5 similar images for each class

are illustrated in Fig. 11. We can clearly find that all the returned images are correct, which also verifies the effectiveness of the proposed method.

Table 4: The mean Average Precision (mAP) for zero shot image retrieval.

Methods	**SUN**	**CUB**	**AWA1**	**aPY**	**Average**
SSE [57]	58.9	4.7	46.25	15.4	31.3
JSLE[58]	76.5	23.9	66.5	32.7	49.9
SynC [6]	74.3	34.3	65.4	30.4	51.1
ISEC [5]	52.7	25.3	68.1	36.9	45.8
MFMR [49]	77.4	30.6	70.8	45.6	56.2
LESD [9]	76.6	31.3	71.2	40.3	54.9
GSDL [10]	79.2	34.2	73.6	44.8	58.0
Ours	**79.5**	**40.7**	**77.2**	**46.1**	**60.9**

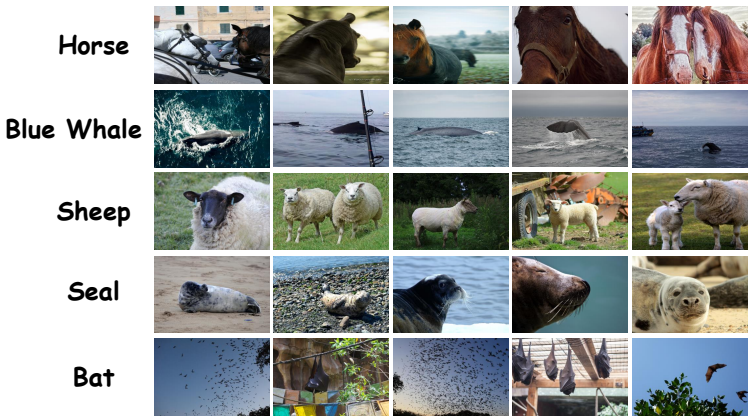


Figure 11: The top-5 retrieval results on AWA.

5. Conclusions

In this paper, we have proposed a novel and effective GZSL image classification model named MIANet, which aims at learning more representative and discriminative latent representations. Specifically, in order to make the

latent vectors more discriminative, we employ a bi-orthogonal constraint in latent hyper-spherical space. In addition, an adversarial training method is conducted in our network to encourage the latent representations to capture more high-level semantic consistency information. Furthermore, in order to get more modality independent information for the latent vectors, we propose a cross reconstruction subnetwork. Finally, we conduct a minimax training mechanism to optimize the discriminator and the generator. Extensive experiments on all five popular datasets are conducted, and the results on both GZSL and ZSL demonstrate the superiority of our method.

6. Acknowledgement

This work was supported in part by National Natural Science Foundation of China (NSFC) under Grants No. 61872187 and No. 61929104, in part by the Medical Research Council (MRC) Innovation Fellowship (UK) under Grant No.MR/S003916/1, and in part by the "111" Program under Grant No.B13022.

References

[1] Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C., 2013. Label-embedding for attribute-based classification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 819–826.

[2] Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C., 2016. Label-embedding for image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 1425–1438.

[3] Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B., 2015. Evaluation of output embeddings for fine-grained image classification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2927–2936.

[4] Atzmon, Y., Chechik, G., 2018. Probabilistic and-or attribute grouping for zero-shot learning. arXiv preprint arXiv:1806.02664 .

[5] Bucher, M., Herbin, S., Jurie, F., 2016. Improving semantic embedding consistency by metric learning for zero-shot classification, in: European Conference on Computer Vision, pp. 6034–6042.

[6] Changpinyo, S., Chao, W.L., Gong, B., Sha, F., 2016. Synthesized classifiers for zero-shot learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 5327–5336.

[7] Chao, W.L., Soravit, C., Gong, B., Sha, F., 2016. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild, in: European Conference on Computer Vision, pp. 52–68.

[8] Chen, L., Zhang, H., Xiao, J., Liu, W., Chang, S.F., 2018. Zero-shot visual recognition using semantics-preserving adversarial embedding networks, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1043–1052.

[9] Ding, Z., Shao, M., Fu, Y., 2017. Low-rank embedded ensemble semantic dictionary for zero-shot learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2050–2058.

[10] Ding, Z., Shao, M., Fu, Y., 2019. Generative zero-shot learning via low-rank embedded semantic dictionary. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 2861–2874.

[11] Farhadi, A., Endres, I., Hoiem, D., Forsyth, D., 2009. Describing objects by their attributes, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1778–1785.

[12] Feng, F., Wang, X., Li, R., 2014. Cross-modal retrieval with correspondence autoencoder, in: ACM Conference on MultiMedia, ACM. pp. 7–16.

[13] Ferrari, V., Zisserman, A., 2008. Learning visual attributes, in: Annual Conference on Neural Information Processing Systems, pp. 433–440.

[14] Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al., 2013. Devise: A deep visual-semantic embedding model, in: Annual Conference on Neural Information Processing Systems, pp. 2121–2129.

[15] Fu, Y., Hospedales, T.M., Xiang, T., Fu, Z., Gong, S., 2014. Transductive multi-view embedding for zero-shot recognition and annotation, in: European Conference on Computer Vision, pp. 584–599.

[16] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Annual Conference on Neural Information Processing Systems, pp. 2672–2680.

[17] Hardoon, D.R., Szedmak, S., Shawe-Taylor, J., 2004. Canonical correlation analysis: An overview with application to learning methods. Neural computation 16, 2639–2664.

[18] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

[19] Hotelling, H., 1992. Relations between two sets of variates, in: Breakthroughs in statistics, pp. 162–190.

[20] Huang, H., Wang, C., Yu, P.S., Wang, C.D., 2019. Generative dual adversarial network for generalized zero-shot learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 801–810.

[21] Jiang, H., Wang, R., Shan, S., Chen, X., 2018. Learning class prototypes via structure alignment for zero-shot recognition, in: European Conference on Computer Vision, pp. 118–134.

[22] Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes, in: International Conference on Learning Representation.

[23] Kodirov, E., Xiang, T., Gong, S., 2017. Semantic autoencoder for zero-shot learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3174–3183.

[24] Kumar Verma, V., Arora, G., Mishra, A., Rai, P., 2018. Generalized zero-shot learning via synthesized examples, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 4281–4289.

[25] Lampert, C., Nickisch, H., Harmeling, S., 2014. Attribute based classification for zero-shot visual object categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 36, 453–465.

[26] Lampert, C.H., Nickisch, H., Harmeling, S., 2009. Learning to detect unseen object classes by between-class attribute transfer, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 951–958.

[27] Li, D., Dimitrova, N., Li, M., Sethi, I.K., 2003. Multimedia content processing through cross-modal association, in: ACM Conference on MultiMedia, ACM. pp. 604–611.

[28] Liu, S., Long, M., Wang, J., Jordan, M.I., 2018a. Generalized zero-shot learning with deep calibration network, in: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, pp. 2005–2015.

[29] Liu, Y., Gao, Q., Li, J., Han, J., Shao, L., 2018b. Zero shot learning via low-rank embedded semantic autoencoder, in: International Joint Conferences on Artificial Intelligence, pp. 2490–2496.

[30] Long, T., Xu, X., Li, Y., Shen, F., Song, J., Shen, H.T., 2018. Pseudo transfer with marginalized corrupted attribute for zero-shot learning, in: ACM Conference on MultiMedia, pp. 1802–1810.

[31] Long, Y., Liu, L., Shao, L., Shen, F., Ding, G., Han, J., 2017. From zero-shot learning to conventional supervised classification: Unseen visual data

synthesis, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1627–1636.

[32] Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-sne. Journal of machine learning research 9, 2579–2605.

[33] Mishra, A., Krishna Reddy, S., Mittal, A., Murthy, H.A., 2018. A generative model for zero shot learning using conditional variational autoencoders, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshop, pp. 2188–2196.

[34] Ni, J., Zhang, S., Xie, H., 2019. Dual adversarial semantics-consistent network for generalized zero-shot learning, in: Advances in Neural Information Processing Systems, pp. 6146–6157.

[35] Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G.S., Dean, J., 2014. Zero-shot learning by convex combination of semantic embeddings, in: International Conference on Learning Representation.

[36] Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M., 2009. Zero-shot learning with semantic output codes, in: Annual Conference on Neural Information Processing Systems, pp. 1410–1418.

[37] Patterson, G., Xu, C., Su, H., Hays, J., 2014. The sun attribute database: Beyond categories for deeper scene understanding. International Journal of Computer Vision 108, 59–81.

[38] Peng, Y., Huang, X., Qi, J., 2016. Cross-media shared representation by hierarchical learning with multiple deep networks., in: International Joint Conferences on Artificial Intelligence, pp. 3846–3853.

[39] Romera-Paredes, B., Torr, P., 2015. An embarrassingly simple approach to zero-shot learning, in: ICML, pp. 2152–2161.

[40] Schonfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., Akata, Z., 2019. Generalized zero-and few-shot learning via aligned variational autoencoders, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 8247–8255.

[41] Socher, R., Ganjoo, M., Manning, C.D., Ng, A., 2013. Zero-shot learning through cross-modal transfer, in: Annual Conference on Neural Information Processing Systems, pp. 935–943.

[42] Verma, V.K., Rai, P., 2017. A simple exponential family framework for zero-shot learning, in: the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Springer. pp. 792–808.

[43] Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001. California Institute of Technology.

[44] Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T., 2017. Adversarial cross-modal retrieval, in: ACM Conference on MultiMedia, pp. 154–162.

[45] Wang, W., Pu, Y., Verma, V.K., Fan, K., Zhang, Y., Chen, C., Rai, P., Carin, L., 2018. Zero-shot learning via class-conditioned deep generative models, in: Thirty-Second AAAI Conference on Artificial Intelligence, pp. 4211–4218.

[46] Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., Schiele, B., 2016. Latent embeddings for zero-shot classification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 69–77.

[47] Xian, Y., Lorenz, T., Schiele, B., Akata, Z., 2018. Feature generating networks for zero-shot learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

[48] Xian, Y., Schiele, B., Akata, Z., 2017. Zero-shot learning-the good, the bad and the ugly, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 4582–4591.

[49] Xu, X., Shen, F., Yang, Y., Zhang, D., Tao Shen, H., Song, J., 2017. Matrix tri-factorization with manifold regularizations for zero-shot learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

[50] Zhang, F., Shi, G., 2019. Co-representation network for generalized zero-shot learning, in: Proceedings of the 36th International Conference on Machine Learning, pp. 7434–7443.

[51] Zhang, H., Koniusz, P., 2018. Zero-shot kernel learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 7670–7679.

[52] Zhang, H., Long, Y., Guan, Y., Shao, L., 2019a. Triple verification network for generalized zero-shot learning. IEEE Transactions on Image Processing 28, 506–517.

[53] Zhang, H., Long, Y., Liu, L., Shao, L., 2019b. Adversarial unseen visual feature synthesis for zero-shot learning. Neurocomputing 329, 12–20.

[54] Zhang, H., Long, Y., Yang, W., Shao, L., 2019c. Dual-verification network for zero-shot learning. Information Sciences 470, 43–57.

[55] Zhang, W., Zuo, Z., Wang, Y., Zhang, Z., 2019d. Double-integrator dynamics for multiagent systems with antagonistic reciprocity. IEEE Transactions on Cybernetics doi:10.1109/TCYB.2019.2939487.

[56] Zhang, Y., Liu, Y., 2020. Nonlinear second-order multi-agent systems subject to antagonistic interactions without velocity constraints. Applied Mathematics and Computation 364. doi:10.1016/j.amc.2019.124667.

[57] Zhang, Z., Saligrama, V., 2015. Zero-shot learning via semantic similarity embedding, in: Internation Conference on Computer Vision, pp. 4166–4174.

[58] Zhang, Z., Saligrama, V., 2016. Zero-shot learning via joint latent similarity embedding, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 6034–6042.