Attentive Semantic Preservation Network for Zero-Shot Learning

Ziqian Lu*
Zhejiang University
ziqianlu@zju.edu.cn

Yunlong Yu*
Zhejiang University
yuyunlong@zju.edu.cn

Zhe-Ming Lu†
Zhejiang University
zheminglu@zju.edu.cn

Feng-Li Shen
Zhejiang University
fenglishen@zju.edu.cn

Zhongfei Zhang
Binghamton University
zhongfei@cs.binghamton.edu

Abstract

While promising progress has been achieved in the Zero-Shot Learning (ZSL) task , the existing generated approaches still suffer from overly plain pseudo features, resulting in poor discrimination of the generated visual features. To improve the quality of the generated features, we propose a novel Attentive Semantic Preservation Network (ASPN) to encode more discriminative as well as semantic-related information into the generated features with the category self-attention cues. Specifically, the feature generation and the semantic inference modules are formulated into a unified process to promote each other, which can effectively align the cross-modality semantic relation. The category attentive strategy encourages model to focus more on intrinsic information of the noisy generated features to alleviate the confusion of generated features. Moreover, prototype-based classification mechanism is introduced in an efficient way of leveraging known semantic information to further boost discriminative of the generated features. Experiments on four popular benchmarks, i.e., AWA1, AWA2, CUB, and FLO verify that our proposed approach outperforms state-of-the-art methods with obvious improvements under both the Traditional ZSL (TZSL) and the Generalized ZSL (GZSL) settings.

1. Introduction

Rcently, driven by the big data, the deep learning models have been achieving great success in the field of computer vision, especially for the image classification task. Although the deep learning models perform well on the traditional supervised classification task, they will fail to classify novel objects that no or few visual samples are available. In reality, it is impractical to collect or annotate enough data for some rare categories. Inspired by this, the researchers try to endow the learning system with the ability to recognize unseen classes via exploiting their auxiliary semantic information, such a task is called Zero-Shot Learning (ZSL).

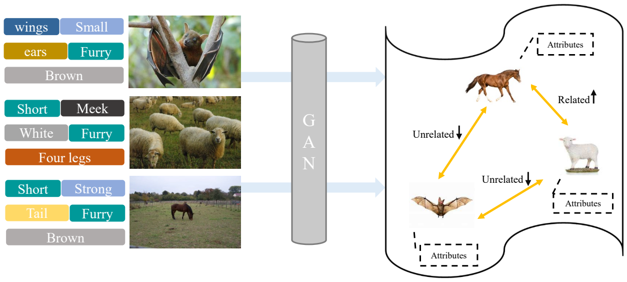


Figure 1. ASPN generates pseudo features and infers the corresponding semantic attributes simultaneously. The inter-class relation between generated features can be calculated attentively as the category self-attention cues to further guide the generation process.

The existing ZSL approaches [12, 17, 13, 21] typically tackle this task by leveraging the seen classes to learn interaction models to connect the visual instances and their corresponding class semantic information, which are then applied to unseen classes. In addition, benefited from the success of generative adversarial networks (GANs) [7, 18], the generated approaches are predominant to address ZSL task. The current generated approaches mostly are based on conditioned GAN [15] that generates visual features for unseen classes by taking both the class semantic features and noises as inputs. Then, generated visual features are used to train a standard classifier for object recognition.

Despite the promising results of these methods, the existing generated approaches mostly suffer from the following issues. First, most of the previous approaches only establish a mapping between semantic and visual in a single direction, which does not guarantee that the generated visual fea-

[*] These authors contributed equally to this work.
[†] Corresponding author.

tures are semantic-related. Second, more attention should be paid to the discriminative of generated features because the generated visual features are finally used for classification in the ZSL task. To address the above problems, we propose a novel generated model named Attentive Semantic Preservation Network (ASPN) to improve the quality of the generated features as illustrated in Fig. 1. Specifically, the proposed model consists of three modules, a semantic preservation network, a category attentive network, and a prototype-based classification network. The semantic preservation network formulates both the feature generation and semantic inference in an cross-modality manner to achieve visual and semantic bidirectional alignment. The category attentive network aims at capturing interclass relation to further improve the discriminative power of the model and alleviate confusion between similar categories, where the attention layer is added proportionally to the generated visual space to gradually guide the generation of higher quality visual features. Besides, prototype-based classification branch is introduced by optimizing the decision boundary at the semantic level to avoid introducing extra parameters, which is more efficient than training a classifier for classification. With the mutual promotion of these three modules, the proposed model generates more discriminative and semantic-related features through semantic constraints with category attention cues and class-level discriminative information. In summary, the contributions of this paper are:

- We propose a novel Attentive Semantic Preservation approach that takes advantage of semantic constraints to formulate both the feature generation and semantic inference simultaneously to achieve visual and semantic bidirectional alignment.

- To further improve the quality of the generated features, we propose a novel category attentive module and a prototype-based classification module to separately capture intrinsic and discriminative information from genered features.

- We conduct extensive experiments on four benchmark datasets, i.e., AWA1, AWA2, CUB, and FLO on both ZSL and GZSL settings. Our model establishes the new state-of-the-art performance in both settings on these datasets, especially achieves remarkable performances on the challenging fine-grained datasets.

2. Related Work

2.1. Generated Models for ZSL

In recent years, significant progress in the generated approaches suggests yielding the desired distribution with a simple instance via functional approximators. Motivated by this idea, some models are proposed to generate pseudo features for unseen classes with adversarial networks [27, 30] and variational autoencoder [25]. CLSWGAN [27] and GAZSL [30] try to generate unseen samples and train a classifier for them. LisGAN [14] regularizes that each sample should be close to one invariant prototype sample. GDAN [9] regularizes an extra regressor with dual learning, which is related to our work. We instead introduce prototype-based classification in cross-modality form to improve the discriminative of generation. Another related work Cycle-CLSWGAN [5] proposes to use the cycle feedback loss as constraints. Different from Cycle-CLSWGAN [5], our model combines the feature generation and the semantic inference to ensure that the generated features are more semantic-related and pay more attention to the intrinsic relation between categories.

3. Attentive Semantic Preservation Model

The overall framework of Attentive Semantic Preservation Network is illustrated in Fig. 2, where the model consists of three core components, a semantic preservation generated module, a category attentive module and a prototype-based classification module. In this section, we first formally give a formulation overview of our proposed model, and then introduce each part of our model in details.

3.1. Problem Definition and Notations

In ZSL and GZSL the training data is denoted by set $\mathcal{S} = \left\{(\mathbf{x}_i, \mathbf{s}_i, \mathbf{y}_i)\,|\mathbf{x}_i \in \mathcal{X}^{\mathcal{S}}, \mathbf{y}_i \in \mathcal{Y}^{\mathcal{S}}, \mathbf{s}_i \in \mathcal{R}^q\right\}$, where $\mathbf{x}_i$ represents images feature of set $\mathcal{X}^{\mathcal{S}}$ extracted by convolutional neural network for seen classes, $\mathbf{y}_i$ represents label of set $\mathcal{Y}^{\mathcal{S}}$ for each seen classes, $\mathbf{s}_i$ represents semantic vector (attribute or sentence description) with q dimensions for each seen class i. Then in test stage, we are given similar test set $\mathcal{U} = \left\{(\mathbf{x}_t, \mathbf{s}_t, \mathbf{y}_t)\,|\mathbf{x}_t \in \mathcal{X}^{\mathcal{U}}, \mathbf{y}_t \in \mathcal{Y}^{\mathcal{U}}, \mathbf{s}_t \in \mathcal{R}^q\right\}$, where $\mathcal{X}^{\mathcal{U}}$ is the set of visual features from unseen classes, $\mathcal{Y}^{\mathcal{U}}$ is the set of unseen classes labels and $\mathcal{Y}^{\mathcal{S}} \cap \mathcal{Y}^{\mathcal{U}} = \emptyset$, $\mathbf{s}_t$ represents semantic vector of unseen classes. Formally, the goal of ZSL is to learn a classifier $f_{zsl} : \mathcal{X}^{\mathcal{U}} \rightarrow \mathcal{Y}^{\mathcal{U}}$. Similarly, we learn a classifier for GZSL $f_{gzsl} : \mathcal{X}^{\mathcal{S}}, \mathcal{X}^{\mathcal{U}} \rightarrow \mathcal{Y}^{\mathcal{S}} \cup \mathcal{Y}^{\mathcal{U}}$.

3.2. Semantic Preservation Network

Given the set $\left\{\mathbf{x}_i \in \mathcal{X}^{\mathcal{S}}, \mathbf{y}_i \in \mathcal{Y}^{\mathcal{S}}\right\}$, the semantic vector of seen samples $\mathbf{s}_i \in \mathcal{R}^q$ and noises $\mathbf{z} \sim N(0, 1)$, the encoder leverages the input $\mathbf{s}_i$ and noises $\mathbf{z}$ to generate pseudo visual features. Thus, during the training, the pseudo features are supervised with real visual features:

$$\mathcal{L}_{\mathcal{V}} = \min_{\theta} \sum_{i} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2\,, \tag{1}$$

where $\hat{\mathbf{x}}_i = E_\theta(\mathbf{s}_i, \mathbf{z})$ is the generated features with semantic vector $\mathbf{s}_i$ and random noises $\mathbf{z}$; θ is the parameter

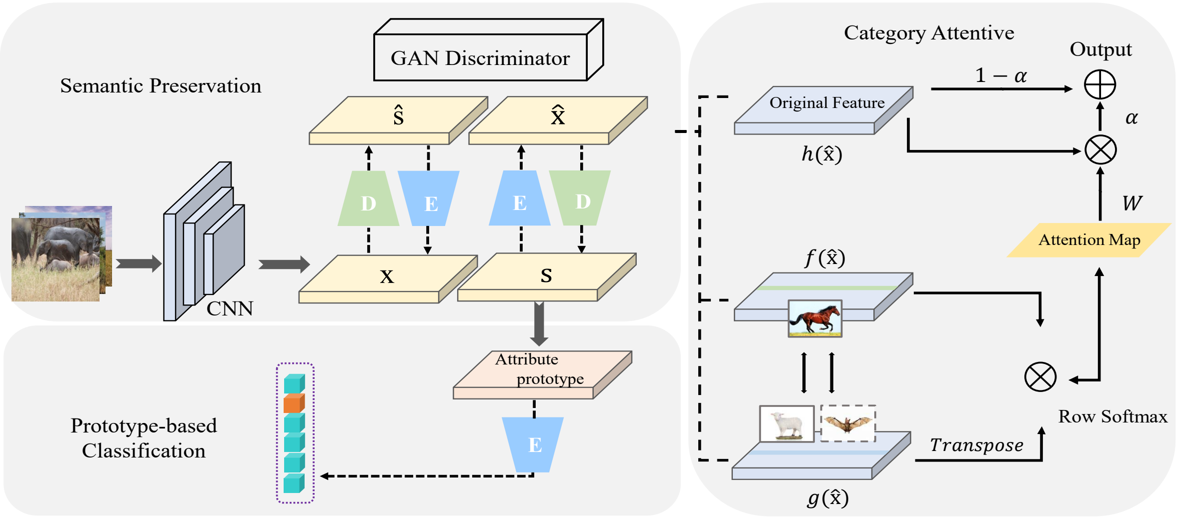


Figure 2. Architecture illustration of our model, including Semantic Preservation (SP), Category Attentive (CA) and Prototype Classification (PC) modules. The visual and semantic features are represented by x and s, respectively. The generated features x̂ given by SP are semantically rich and boundary-distinguishable with the help of the PC. The category attention obtained by matrix multiplication and row softmax is proportionally applied to the original generated features from SP to produce the final visual features.

of the encoder. This term encourages that the generated visual features are similar to the real visual features. At the same time, the decoder takes training samples as input to decompose the input into pseudo attributes and noise vectors, where pseudo attribute vectors are supervised with real semantic information. It should be noted that we write $\hat{\mathbf{s}}_i, \hat{\mathbf{z}} = D_\mu(\mathbf{x}_i)$, where μ is the parameter of the decoder and $\hat{\mathbf{z}}$ are the generated noise vectors. Intuitively, the generated class semantic vector $\hat{\mathbf{s}}_i$ should be close to the real class semantic prototype, i.e.,

$$\mathcal{L}_{\mathcal{S}} = \min_{\theta,\mu} \sum_i \|\mathbf{s}_i - \hat{\mathbf{s}}_i\|_2^2 , \tag{2}$$

Inspired by the effectiveness of the modality consistent, we introduce the unified structure for generated features to reduce the difference for both visual generation space and semantic inference space, respectively. Specifically, we preserve the visual and semantic prototype by reusing the pretrained E_θ and D_μ to achieve alignment for each space. The training objective of semantic preservation part is to minimize the following loss function:

$$\begin{aligned}\mathcal{L}_{pre}(\theta,\mu) = \mathbb{E}_{\mathbf{s}\sim p(\mathbf{s})}&\left[\|\mathbf{s}_i - D_\mu(E_\theta(\mathbf{s}_i,\mathbf{z}))\|_2^2\right] \\ +\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}&\left[\|\mathbf{x}_i - E_\theta(D_\mu(\mathbf{x}_i))\|_2^2\right]\end{aligned} \tag{3}$$

With Eq. (3) the feature generation and semantic inference procedures are formed as more powerful unified constraint to encourage generated visual features from the same class to be clustered together while preserving distribution and semantic-related information.

As the visual and class semantic features are high-level representations, the l_p- norm is hard to capture sufficient information. Hence, we also adopt adversarial learning for the visual semantic pairs. The discriminator is designed to distinguish whether the input is from the output of the generator or the real data distribution. The adversarial process is formulated as:

$$\begin{aligned}\mathcal{L}_D =& \mathbb{E}_{(\mathbf{x},\mathbf{s})\sim p(\mathbf{x},\mathbf{s})}\left[D_\varphi(\mathbf{x},\mathbf{s})\right] \\ &- \mathbb{E}_{(\hat{\mathbf{x}},\mathbf{s})\sim p_\theta(\hat{\mathbf{x}}|\mathbf{s},\mathbf{z})}\left[D_\varphi(\hat{\mathbf{x}},\mathbf{s})\right] - \beta\mathcal{L}_{GP},\end{aligned} \tag{4}$$

where D is discriminator with parameter φ and $\mathcal{L}_{GP} = (\|\nabla_\tau D_\varphi(\tau,\mathbf{s})\|_2^2 - 1)^2$ is the gradient penalty to enforce the Lipschitz constraint; τ is the linear interpolation between the real feature $\mathbf{x}$ and the generated feature $\hat{\mathbf{x}}$; $\beta > 0$ is a hyper parameter.

3.3. Category Attentive Network

The above Semantic Preservation Network mostly focuses on learning to generate more semantic-related visual features. However, it is no doubt that the feature confusion does exist in traditional and generalized zero-shot learning, especially generated methods. In this section we introduce a Category Attentive Network that takes the task of guiding generator to generate more discriminative and high quality visual features. As illustrated in the right of Fig. 2, the generated features from the previous hidden layer $\hat{\mathbf{x}} \in \mathcal{R}^{b\times d}$ are first copied into two feature spaces $f(\hat{\mathbf{x}}) \in \mathcal{R}^{b\times d}$ and $g(\hat{\mathbf{x}}) \in \mathcal{R}^{b\times d}$ to calculate the attention and $h(\hat{\mathbf{x}}) \in \mathcal{R}^{b\times d}$ represents original features, where b is batch size of the generated samples and d is the number of feature dimensions. Here, we have Eq. (5):

$$k_{ij} = f(\hat{\mathbf{x}}_i)\, g(\hat{\mathbf{x}}_j)^T, W_{i,j} = \frac{\exp(k_{ij})}{\sum_{i=1}^b \exp(k_{ij})}, \tag{5}$$

where the i^{th} row and j^{th} column of k reflects the influence to which the j^{th} class attends to the i^{th} class, in other words, the product between the different generated features is regarded as a category correlation matrix. Hence, each row of W after softmax represents a mode of attention and the output is formulated as:

$$\mathbf{t}=\sum_{i=1}^{b} W_{i,j} h\left(\hat{\mathbf{x}}\right), \mathbf{o}_i=\alpha \mathbf{t}+(1-\alpha) h(\hat{\mathbf{x}}), \qquad (6)$$

where W is attention map, the original features are multiplied by the attention weights to produce the category attentive representation $\mathbf{t}$, and α is a parameter to control the influence of attention for final generated features $\mathbf{o}$ based on the consideration of generated feature quality and iterative stability. More intuitively, assuming that the i^{th} row of f is a horse and the j^{th} column of g^T is a sheep, the correlation between the two is relatively strong, and the correlation is weak if the j^{th} column represents a bat. Then we can learn better features through attention and it helps to refine the category feature, enabling the similar images to have smaller distances and dissimilar images have larger distances. Besides, the category attention is obtained from the previous generated process as the intrinsic information. This indicates that this strategy has potential for expanding into other methods and tasks. In this way, the attention layer is seen as the auxiliary generation module, which allows the network to first rely on the cues from a certain category and then gradually learn to assign more weight to the inter-class evidence. This also mitigates the classification bias towards some specific classes. Therefore, the generated features are more high-quality and discriminative.

3.4. Prototype-based Classification Network

Since the goal of ZSL and GZSL is to identify novel classes, the quality and generalization ability of generated features should be focused. To further preserve the discriminative information of generated visual features, a classification network is designed to predict corresponding class labels of both seen and unseen categories. We introduce the classification branch from the perspective of prototype to reduce model complexity and make full use of known semantic information. In fact, the experiments show that the proposed strategy is more reasonable. The training objective of classification part is to minimize the following loss function:

$$\mathcal{L}_{cls}=\min_{\theta} \sum_{i}\left(L_{\theta}\left(\mathbf{x}_i, \mathbf{A}\right)+L_{\theta}\left(\hat{\mathbf{x}}_i, \mathbf{A}\right)\right), \qquad (7)$$

$$L_{\theta}\left(\mathbf{x}_i, \mathbf{A}\right)=-\log P\left(\mathbf{y} \mid \mathbf{x}_i, \mathbf{A} ; \theta\right), \qquad (8)$$

where $P\left(\mathbf{y}_j \mid \mathbf{x}_i, \mathbf{A} ; \theta\right)=\frac{\exp\left(\mathbf{x}_i^T E_{\theta}(\mathbf{s}_j)\right)}{\sum_k^M \exp\left(\mathbf{x}^T E_{\theta}(\mathbf{s}_j)\right)}$, and L_{θ} are classification losses of real and generated pseudo visual features, $\mathbf{A} \in \mathcal{R}^{q \times M}$ is the class semantic prototype matrix of both the seen and unseen classes, M is number of all classes with semantic features dimensions q. While, $\mathbf{s}_j$ is the corresponding class semantic prototype of class $\mathbf{y}_j$, E_{θ} mentioned in Eq. (3) is used to project the class semantic features into the visual space. The value of $\mathbf{x}_i^T E_{\theta}\left(\mathbf{s}_j\right)$ is seen as the compatibility score between the visual feature $\mathbf{x}_i$ and the j^{th} class semantic prototype $\mathbf{s}_j$ selected from semantic space. If the sample $\mathbf{x}_i$ belongs to class $\mathbf{y}_j$, their compatibility score should be large; otherwise it should be small. In this way, the separability between any two different classes is enlarged at the level of visual-semantic prototypes. Besides, the unseen class semantic prototypes are also taken into consideration, which prevents the seen data from classifying into unseen classes. The seen to unseen bias issue thus is mitigated obviously.

3.5. Full Objective

Overall, the objective function trained of the proposed model is summarized with:

$$\mathcal{L}_{Obj}=\mathcal{L}_{\mathcal{V}}+\mathcal{L}_{\mathcal{S}}+\mathcal{L}_{pre}+\mathcal{L}_{D}+\vartheta \mathcal{L}_{cls} \qquad (9)$$

where ϑ is hyper-parameter that assign weight on the classification loss. Among them, $\mathcal{L}_{pre}$ and $\mathcal{L}_{cls}$ reuse the corresponding network, so no additional parameters are introduced. During the test stage, the similarities between the test instances and the unseen class semantics prototypes are obtained by calculating the distances of the visual features and the generated unseen visual features.

4. Experiments

In this section, we conduct experiments to evaluate our approach on both traditional ZSL and generalized ZSL. We first document the datasets and the experimental settings and then compare our approach with some selected competitors. Finally, some ablation studies are given, as well as the discussions.

4.1. Datasets and Implementation Details

We conduct experiments on two coarse-grained datasets and two fine-grained datasets. Animals with Attributes1 (AWA1) [11] consists of 30,475 images of 50 animal species, where 85-dimensional attributes are provided for each class as the class semantic features. Similarly, AWA2 [22] has 37,322 images with 85-dimensional attributes for 50 classes. In addition to the coarse-grained datasets AWA1 and AWA2, we also test fine-grained datasets Caltech-UCSD-Birds 200-2011 (CUB) [24] and Oxford Flowers (FLO) [16] with 11,788 and 8,189 images, respectively. As for the class semantic for both CUB and FLO datasets, we follow Cycle-CLSWGAN [5] and leverage 1,024-dimensional semantic features produced by the character-based CNN-RNN [19] that encodes the textual description

Method	AWA1	AWA2	CUB	FLO	AWA1			AWA2			CUB			FLO		
	T				**u**	**s**	**H**	**u**	**s**	**H**	**u**	**s**	**H**	**u**	**s**	**H**
LATEM[26]	55.1	55.8	49.3	40.4	7.3	71.7	13.3	11.5	77.3	20.0	15.2	57.3	24.0	6.6	47.6	11.5
DEVISE[6]	54.2	59.7	52.0	45.9	13.4	68.7	22.4	17.1	74.7	27.8	23.8	53.0	32.8	9.9	44.2	16.2
ESZSL[20]	58.2	58.6	53.9	51.0	2.4	70.1	4.6	5.9	77.8	11.0	12.6	63.8	21.0	11.4	56.8	19.0
SAE[10]	53.0	54.1	33.0	-	1.8	77.1	3.5	1.1	82.2	2.2	8.8	18.0	11.8	-	-	-
ALE[1]	59.9	62.5	54.9	48.5	16.8	76.1	27.5	14.0	81.8	23.9	23.7	62.8	34.4	13.3	61.6	21.9
SYNC[4]	54.0	46.6	55.6	-	8.9	87.3	16.2	10.0	90.5	18.0	11.5	**70.9**	19.8	-	-	-
SJE[2]	65.6	61.9	53.9	53.4	11.3	74.6	19.6	8.0	73.9	14.4	23.5	59.2	33.6	13.9	47.6	21.5
DEM[29]	68.4	67.1	51.7	70.2	32.8	84.7	47.3	30.5	86.4	45.1	19.6	57.9	29.2	57.2	67.7	62.0
RELATION NET[23]	68.2	64.2	55.6	-	31.4	**91.3**	46.7	30.0	**93.4**	45.3	38.1	61.1	47.0	38.9	84.4	53.2
GAZSL[30]	68.2	70.2	55.8	60.5	29.6	84.2	43.8	35.4	86.9	50.3	31.7	61.3	41.8	-	-	-
CLSWGAN[27]	68.2	65.3	57.3	67.2	57.9	61.4	59.6	**56.1**	65.0	60.2	50.3	58.3	54.0	59.0	73.9	65.6
Cycle-CLSWGAN[5]	66.8	-	58.6	70.3	56.9	64.0	60.2	-	-	-	45.7	61.0	52.3	59.2	72.5	65.1
GDAN[9]	-	-	-	-	-	-	-	32.1	67.5	43.5	39.3	66.7	49.5	-	-	-
COSMO[3]	-	-	-	-	52.8	80.0	63.6	-	-	-	44.4	57.8	50.2	59.6	81.4	68.8
LisGAN[14]	70.6	-	58.8	69.9	52.6	76.3	62.3	51.1	72.4	60.0	46.5	57.9	51.6	57.7	83.8	68.3
f-VAEGAN-D2[28]	-	71.1	61.0	67.7	-	-	-	57.6	70.6	**63.5**	48.4	60.1	53.6	56.8	74.9	64.6
ASPN (Ours)	**75.4**	**73.3**	**73.4**	**78.1**	**58.0**	85.7	**69.2**	46.2	87.0	60.4	**50.7**	61.5	**55.6**	**67.3**	**87.4**	**76.0**

Table 1. The per-class average Top-1 accuracy (%) for the traditional (TZSL) and generalized ZSL (GZSL) on four datasets. The best results are marked with boldface.

of an image containing fine-grained visual descriptions (10 sentences per image). As for visual features, we follow the previous work [27] to use the 2,048-dimensional top pooling units of the 101-layered ResNet [8] as the deep features without fine-tuning.

In our model, the encoder and decoder are implemented by three-layer neural network of 1,800 hidden units with ReLU activation. The proposed model is trained with batch size of 64 to find the best one model for testing accuracy. As for parameters, we set $\vartheta = 1.0$, $\beta = 0.01$ and initialize $\alpha = 0.9$. The model is optimized with the Adam solver with a cross-validated learning rate 0.0001. For the traditional ZSL task that all the test instances are from the unseen classes where they are predicted into, we use the average per-class top-1 accuracy **T** to evaluate the performance of each model. For the generalized ZSL task, we evaluate the the accuracy of unseen and seen classes and their harmonic mean to comprehensively show the performance of the proposed model. In harmonic mean **H**=(2×**u**×**s**)/(**u**+**s**), **u** and **s** denote the accuracy of unseen and seen classes, respectively.

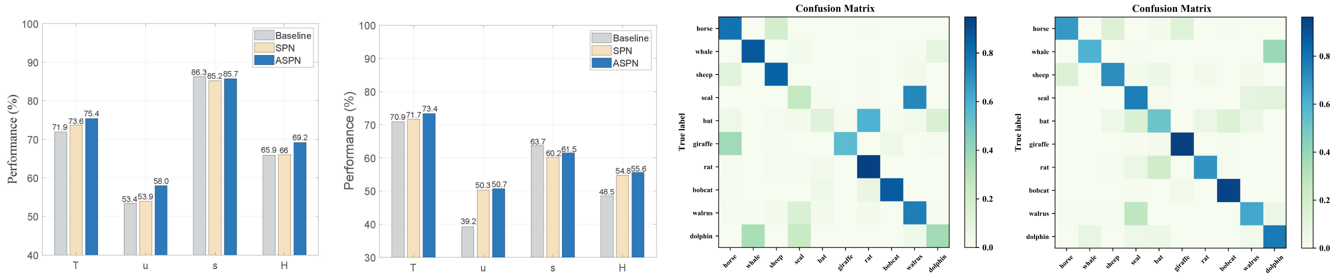
4.1.1 Results on Traditional ZSL

We report the comparison results of the proposed approach and several competitors in Table 1 **T** column. It can be observed that our model achieves the best performance on four datasets. The overall improvements on AWA1, AWA2, CUB, and FLO datasets are from 70.6%, 71.1%, 61.0% and 70.3% to 75.4%, 73.3%, 73.4% and 78.1% i.e., all quite significant, against the previous state-of-the-art. From a holistic perspective, ASPN obtains 75.1% average performance from coarse-grained to fine-grained datasets. It should be noted that our approach achieves excellent results on two challenging fine-grained datasets CUB and FLO. This indicates that the proposed unified model is able to accurately distinguish a large number of different categories through the SP and CA strategies. The contribution of each part will be discussed in the ablation study.

4.1.2 Results on Generalized ZSL

We then report the results of the generalized ZSL task and compare the proposed model with these approaches in Table 1.

From the result, it can be observed that our approach achieves superior and stable results on most datasets under generalized setting, especially on **T** and **H** metric. Specifically, for the coarse dataset AWA1, we obtain 5.6% improvement in terms of comprehensive evaluation metric **H** over the second best approach COSMO[3], which indicates that our ASPN achieves balance between the seen and unseen accuracy benefited from the higher quality generated samples. Compared with similar cyclic based work Cycle-CLSWGAN [5], ASPN greatly improves the performance on unseen accuracy, which proves that the category attentive regularization may further promote the diversity and discrimination of the genrated features. As a result, it can tackle the confusing problem of seen and unseen classes to boost the harmonic mean accuracy. On the AWA2 dataset, our model is still very competitive, where the proposed model obtains the highest **H**. Different from the best results of f-VAEGAN-D2[28], our ASPN improves the accuracy of unseen classes as much as possible without sacrificing the accuracy of the seen class, which is more feasible in real life. In summary, our method generates discriminative features by the proposed strategies to alleviate part of the classification bias problems and balance the accuracy of seen and unseen categories under the challenging GZSL setting to boost the performance significantly.



(a) AWA1 (b) CUB (c) CLSWGAN (d) ASPN

Figure 3. The classification results (in %) of the traditional and generalized ZSL for component analysis in (a) AWA1, (b) CUB and the confusion matrices on the evaluation of AWA1 dataset for both CLSWGAN and ASPN (Ours) method.

Model	MFLOPs	Parameters Memory
Cycle-CLSWGAN	304.3	18.1M
ASPN (Ours)	109.8	13.9M

Table 2. Complexity evaluation with FLOPs and Parameters Memory under Tensorflow framework.

4.2. Ablation Studies

In this section, we analyze our Attentive Semantic Preservation Network in terms of components of the model and give the class-wise accuracy evaluation.

4.2.1 Component Analysis.

In order to investigate the impacts of each module, we conduct the component analysis on these datasets. Specifically, the unidirectional mapping of the generation from semantic space with prototype-based classification loss is set to be the **Baseline**. At the same time, **SPN** represents semantic preservation module and **ASPN** means the complete network equipped with all the components. As can be seen from Fig. 3 (a) (b), the proposed model with **SPN** module outperforms the baseline across all these datasets in terms of unseen accuracy, where CUB obtains the highest 11.1% improvement for **u**. Besides, our ASPN method further improves the performance of unseen classes while maintaining the accuracy of seen classes. In fact, the results (i.e.,**T** and **u**) of the **ASPN** model with category self-attention are boosted compared to the **Baseline** and the **SPN** with a large margin. This indicates that the proposed category self-attention module brings positive impacts for the classification via improving the discriminative of the generated visual features. Note that the accuracy of the seen class is slightly reduced because the generated visual features are a little biased towards the unseen classes.

In addition, Table 2 shows that the proposed prototype-based classification model is more efficient with the lower complexity and better performance than the method of training classifiers. Low runtime consumption is also more realistic in application scenarios.

4.2.2 The Evaluation of the Class-wise Accuracy.

To show the results in detail, we also conduct experiments to evaluate the class-wise accuracy. Fig. 3 (c) (d) shows the confusion matrices of CLSWGAN [27] (**T** =68.2%) and our method (**T** =75.4%) on the AWA1 dataset. Comparing the results of ZSL, we can find that our model has fewer misclassifications, and the classification results are more consistent with the visual correspondence. In fact, our model has better performance on most of the classes, which proves the effectiveness of our method on handling categories fusion in ZSL setting. More intuitively, CLSWGAN [27] tends to recognize seal as walrus in Figure (c), while ASPN can accurately distinguish this visually similar animals with the help of category attention. The same phenomenon appears in the classification of bat and rat, giraffe and horse, which further proves the effectiveness of the proposed method. Back to Fig. 3 (a) we can find that the improvement of **u** obviously promotes the overall performance. By the comparison of **u** for different strategies in histogram, we can draw a simple conclusion that the main contribution comes from category attentive strategy.

5. Conclusion

In this paper, we propose a novel generated approach to improve the quality of generated features. Semantic preservation strategy encodes semantic-related information into the generated features. Category self-attention network and prototype-based classification module are then deployed to further alleviate feature confusion in classification. Extensive experiments on four benchmark datasets demonstrate the effectiveness of the proposed model in both traditional ZSL and challenging generalized ZSL tasks. The ablation studies show the impacts of each part of the model. The class-wise evaluation is also given to intuitively explain why the performance of the proposed model can be improved.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *TPAMI*, 38(7):1425–1438, 2016. 5

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015. 5

[3] Yuval Atzmon and Gal Chechik. Adaptive confidence smoothing for generalized zero-shot learning. In *CVPR*, pages 11671–11680, 2019. 5

[4] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, pages 5327–5336, 2016. 5

[5] Rafael Felix, Vijay BG Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, pages 21–37, 2018. 2, 4, 5

[6] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013. 5

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[9] He Huang, Changhu Wang, Philip S Yu, and Chang-Dong Wang. Generative dual adversarial network for generalized zero-shot learning. In *CVPR*, pages 801–810, 2019. 2, 5

[10] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, pages 3174–3183, 2017. 5

[11] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009. 4

[12] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 36(3):453–465, 2014. 1

[13] Angeliki Lazaridou, Elia Bruni, and Marco Baroni. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *ACL*, pages 1403–1414, 2014. 1

[14] Jingjing Li, Mengmeng Jing, Ke Lu, Zhengming Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. In *CVPR*, pages 7402–7411, 2019. 2, 5

[15] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

[16] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. 4

[17] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, pages 1–9, 2014. 1

[18] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1

[19] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, pages 49–58, 2016. 4

[20] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, pages 2152–2161, 2015. 5

[21] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, pages 935–943, 2013. 1

[22] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NIPS*, pages 3483–3491, 2015. 4

[23] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, pages 1199–1208, 2018. 5

[24] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 4

[25] Wenlin Wang, Yunchen Pu, Vinay Kumar Verma, Kai Fan, Yizhe Zhang, Changyou Chen, Piyush Rai, and Lawrence Carin. Zero-shot learning via class-conditioned deep generative models. In *AAAI*, pages 4211–4218, 2018. 2

[26] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, pages 69–77, 2016. 5

[27] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018. 2, 5, 6

[28] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR*, pages 10275–10284, 2019. 5

[29] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, pages 2021–2030, 2017. 5

[30] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, pages 1004–1013, 2018. 2, 5