electronics

MDPI

Article

Generalized Zero-Shot Image Classification via Partially-Shared Multi-Task Representation Learning

Gerui Wang [1,2,*] and Sheng Tang [1,2]

[1] School of Computer Science and Engineering, Central South University, Changsha 410083, China; 204712137@csu.edu.cn
[2] Hunan Engineering Research Center of Machine Vision and Intelligent Medicine, Central South University, Changsha 410083, China
* Correspondence: wjhzxy@csu.edu.cn

Abstract: Generalized Zero-Shot Learning (GZSL) holds significant research importance as it enables the classification of samples from both seen and unseen classes. A prevailing approach for GZSL is learning transferable representations that can generalize well to both seen and unseen classes during testing. This approach encompasses two key concepts: discriminative representations and semantic-relevant representations. "Semantic-relevant" facilitates the transfer of semantic knowledge using pre-defined semantic descriptors, while "discriminative" is crucial for accurate category discrimination. However, these two concepts are arguably inherently conflicting, as semantic descriptors are not specifically designed for image classification. Existing methods often struggle with balancing these two aspects and neglect the conflict between them, leading to suboptimal representation generalization and transferability to unseen classes. To address this issue, we propose a novel partially-shared multi-task representation learning method, termed PS-GZSL, which jointly preserves complementary and sharable knowledge between these two concepts. Specifically, we first propose a novel perspective that treats the learning of discriminative and semantic-relevant representations as optimizing a discrimination task and a visual-semantic alignment task, respectively. Then, to learn more complete and generalizable representations, PS-GZSL explicitly factorizes visual features into task-shared and task-specific representations and introduces two advanced tasks: an instance-level contrastive discrimination task and a relation-based visual-semantic alignment task. Furthermore, PS-GZSL employs Mixture-of-Experts (MoE) with a dropout mechanism to prevent representation degeneration and integrates a conditional GAN (cGAN) to synthesize unseen features for estimating unseen visual features. Extensive experiments and more competitive results on five widely-used GZSL benchmark datasets validate the effectiveness of our PS-GZSL.

Keywords: Generalized Zero-Shot Learning; discriminative; semantic-relevant; image classification; partially-shared multi-task learning; transferable representation

Citation: Wang, G.; Tang, S. Generalized Zero-Shot Image Classification via Partially-Shared Multi-Task Representation Learning. *Electronics* **2023**, *12*, 2085. https://doi.org/10.3390/electronics12092085

Academic Editor: George A. Papakostas

Received: 29 March 2023
Revised: 27 April 2023
Accepted: 29 April 2023
Published: 3 May 2023

1. Introduction

Generalized Zero-Shot Learning (GZSL) [1] has attracted significant research interest due to its ability to transfer knowledge to unseen classes using additional class-level semantic descriptors, such as word vectors [2] or attributes [3]. As an extension of Zero-Shot Learning (ZSL) [3,4], GZSL aims to classify both seen and unseen classes simultaneously during testing. This capability is crucial in various real-world applications where the availability of labeled samples for all possible classes is limited or infeasible [5,6].

A key idea in GZSL is learning transferable representations, which encompass two essential concepts: *discriminative* and *semantic-relevant* features. *Discriminative* features are crucial for accurate category discrimination, possessing strong decision-making power and promoting the classification task of unseen classes. In contrast, *semantic-relevant* facilitates a shared semantic space between seen and unseen classes using pre-defined semantic

descriptors, reflecting the semantic relationships between different classes as accurately as possible. GZSL can be viewed as a multi-task problem, where learning discriminative features optimizes a discrimination sub-task, and learning semantically-relevant features optimizes a visual-semantic alignment sub-task. By adopting a multi-task perspective, GZSL aims to obtain comprehensive representations between tasks that can generalize well to unseen classes during testing. However, since semantic descriptors are not specifically designed for image classification [1,7,8], two main challenges arise: (1) appropriately balancing these sub-tasks and resolving their conflict, and (2) ensuring the stability and expressiveness of learned representations.

Unfortunately, existing methods tend to bypass or ignore these challenges between *discriminative* and *semantic-relevant*, resulting in passable performance on unseen classes. Specifically: (1) some researchers focus solely on semantic-relevant representations through elaborate visual-semantic alignment [8–10], while others concentrate on advanced discrimination techniques to extract more generalizable discriminative representations [11,12]. (2) Furthermore, the conflict between discrimination and visual-semantic alignment is often neglected, as recent methods primarily focus on learning shared representations between these two sub-task [7,13,14]. As a result, their poor generalization can be attributed to the discarding of some task-specific information between sub-tasks, which can be viewed as the "diamond in the rough" for GZSL. Some works in domain generalization (DG) have shown that this specific information could enhance a model's generalization performance when classifying unseen classes [15,16]. For example, in the AWA1 dataset shown in Figure 1. attributes like "Strong, Big" that are not visually discriminative can still reduce the misclassification between tigers and cats. Similarly, visual cues like the ear and nose shape are salient for classifying image samples but not represented in the semantic descriptors.

To address the aforementioned challenges and limitations, we propose a novel partially-shared representation learning network, termed PS-GZSL, which jointly preserves complementary and transferable information between discriminative and semantic-relevant features. First, to resolve the conflict between tasks and avoid information loss, PS-GZSL proposes a partially-shared multi-task learning mechanism to explicitly model both task-shared and task-specific representations. As depicted in Figure 2, PS-GZSL utilizes three Mixture-of-Experts (MoE) [17,18] to factorize a visual feature into three latent representations: a task-shared discriminative and semantic representation h_{ds}, a task-specific discriminative representation h_d, and a task-specific semantic-relevant representation h_s. Each sub-task corresponds to a task-specific and a task-shared representation. Second, to ensure the stability and expressiveness of learned representations, PS-GZSL draws inspiration from the success of contrastive learning [19] and metric learning [20], proposing two effective sub-tasks: an instance-level contrastive discrimination task and a relation-based visual-semantic alignment task. These tasks have been proven to achieve better generalization performance, respectively. To avoid representation degeneration, PS-GZSL randomly drops out experts in each MoE. Furthermore, PS-GZSL is a hybrid GZSL framework that integrates with a feature generation component. In feature generation, PS-GZSL adopts a conditional generative adversarial network [21] with a feedback mechanism to mitigate the bias towards seen classes in the latent representation space.

In summary, the main contributions of our work can be summarized:

1. We describe a novel perspective grounded in multi-task learning, which reveals that existing methods exhibit an inherent generalization weakness of losing some transferable visual features.
2. We propose a novel GZSL method, termed partially-shared multi-task representation learning network (PS-GZSL), to jointly preserve complementary and transferable information between discriminative and semantic-relevant features
3. Extensive experiments on five widely-used GZSL benchmark datasets validate the effectiveness of our PS-GZSL and show that the joint contributions of the task-shared and task-specific representations result in more transferability representation.

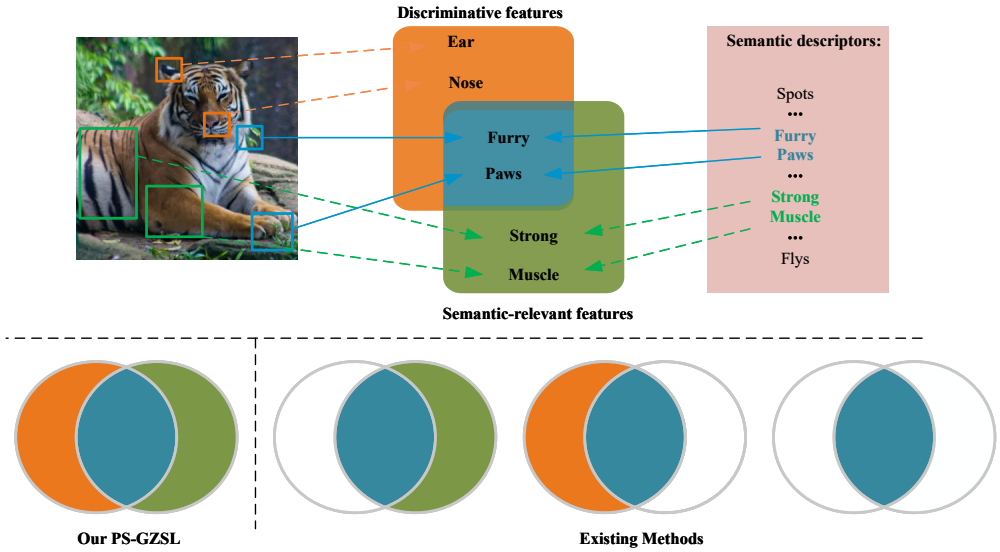


Figure 1. Existing GZSL methods either bypass or ignore the conflict between discriminative and semantic-relevant objectives, and may overlook some task-specific visual features (as indicated by the green and orange dashed lines). In contrast, PS-GZSL can preserve more complete sharable features.

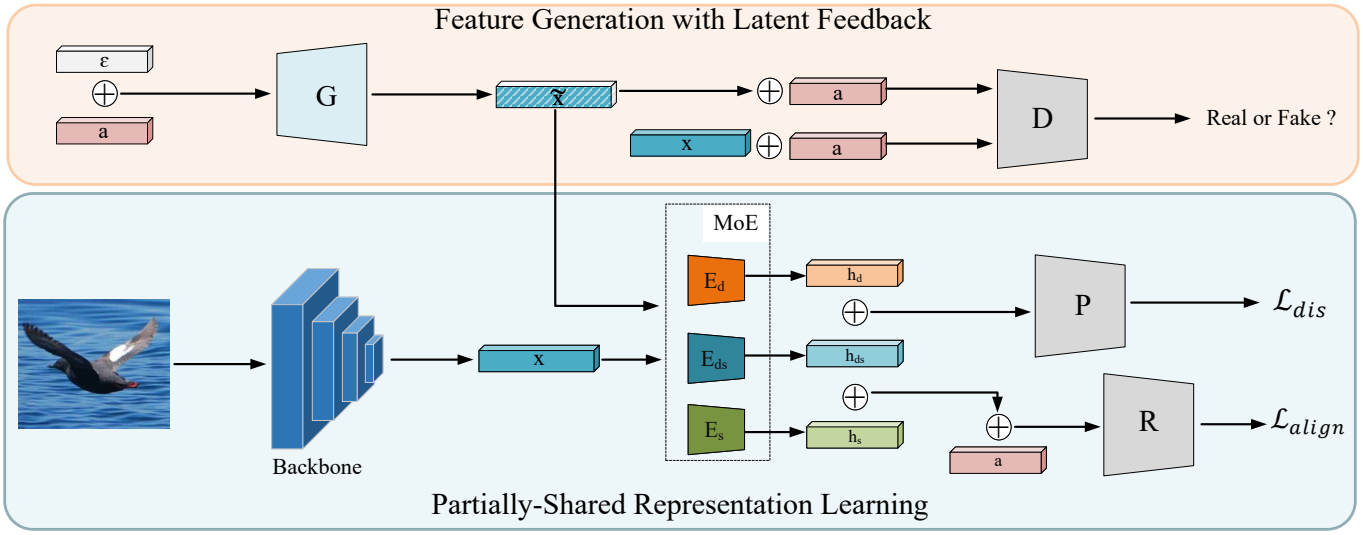


Figure 2. Illustration of our proposed PS-GZSL, which consists of (i) a conditional GAN network **D** and **G** with latent feedback mechanism; (ii) a multi-branch MoE network $E = [E_{ds}, E_d, E_s]$ for factorized latent representation learning. And two task modules **P** and **R** are extended to ensure the discriminative property and semantic property. Here, **a** denotes the semantic descriptors, and ϵ is a random gaussian noise.

2. Related Works

Early approaches for ZSL/GZSL can be broadly classified into two main groups: Embedding-based methods and Generative-based methods. The former group [22–27] learns an encoder to map the visual features of seen classes to their respective semantic descriptors. In contrast, the latter group [21,28–32] learns a conditional generator, such as cVAE [33] or cGAN [34], to synthesize virtual unseen features based on the seen samples and semantic descriptors of both classes.

Recent state-of-the-art methods typically graft an encoder on top of a conditional generator, with a focus on improving the transferability of visual representations. (1) Some methods emphasize preserving semantic-relevant information that corresponds to pre-

defined descriptors. For example, CADA-VAE [9] employs two aligned Variational Autoencoders(VAEs) to learn shared latent representations between semantic descriptors and visual features. SDGZSL [10] integrates a disentanglement constraint and a Relation network [20] to ensure the semantic-consistency of the learned representation. SE-GZSL [35] uses two AutoEncoders and Mutual information maximization to capture semantic-relevant information. (2) Some others prioritize the preservation of more discriminative information. DLFZRL [11] adopts a hierarchical factorizing approach and adversarial learning to learn the discriminative latent representation, regardless of whether it is semantically relevant or not. DR-GZSL [7] utilizes an auxiliary classifier and a shuffling disentanglement mechanism to extract the discriminative part of the semantic-relevant representation. CE-GZSL [13] integrates the semantic-supervised learning module and label-supervised discrimination module in the latent space to learn discriminative visual representations. In summary, these methods differ in the transferable characteristics of the data they model for recognition.

In contrast to existing methods, we argue that both discriminative and semantic-relevant representations are important for recognizing test classes. However, due to the conflict between them, these methods implicitly discard some valuable features. We are thus motivated to adopt the soft-parameter sharing mechanism [17,36] in multi-task learning. This flexibility stems is derived from information routing between tasks, and its characteristics of seeking similarities while preserving differences have led to significant successes in multi-task learning domains such as recommendation systems. We are the first to apply this idea and revise it for representation learning in GZSL. A novel multi-task representation learning paradigm is proposed that models task-specific and task-shared representations in parallel, unlike existing paradigms [37,38] that use a single MoE for each sub-task and a hierarchical structure. For the sake of clear understanding, we highlight the distinctions between our approach and those counterparts in Table 1.

Table 1. Qualitative Model Comparison. The ◯, □, and △ denote representations that are discriminative and semantic-relevant, only discriminative, and only semantic-relevant, respectively.

Model Comparison	Task-Shared	Task-Specific	
	◯	□	△
SP-AEN [8]	√		√
CADA-VAE [9]	√		
SDGZSL [10]	√		√
DLFZRL [11]	√	√	
DR-GZSL [7]	√		
CE-GZSL [13]	√		
Our PS-GZSL	√	√	√

3. Methods

To learn more transferable representations, in this section, we present our proposed PS-GZSL method, which combines MoE, a partially-shared mechanism, an instance contrastive discrimination module, and a relation-based visual-semantic alignment module. To alleviate the bias towards seen, we also adopt a feature generation module with latent feedback. The overall framework of our proposed PS-GZSL is shown in Figure 2, Then, the definition of the ZSL/GZSL problem and all the above modules are explained in detail.

3.1. Problem Definition

In Zero-Shot learning, we are given two disjoint sets of classes: $\{\mathcal{X}^s, \mathcal{Y}^s\}$ with $\mathcal{S}$ seen classes and $\{\mathcal{X}^u, \mathcal{Y}^u\}$ with $\mathcal{U}$ unseen classes, where we have $\mathcal{Y}^s \cap \mathcal{Y}^u = \varnothing$ and $\mathcal{Y}^{all} = \mathcal{Y}^s \cup \mathcal{Y}^u$. For the semantic descriptors $\mathcal{A} = \{a_1, \ldots, a_S, a_{S+1}, \ldots, a_{S+U}\}$, each class, whether seen or unseen, is associated with a semantic descriptor that can take the form of sentences or attributes. Under ZSL setting, we have $\{\mathcal{X}^s, \mathcal{Y}^s, \mathcal{A}^s\}$ and $\{\mathcal{Y}^u, \mathcal{A}^u\}$ available

during training phase. Let $x \in \mathcal{X}$ denote the extracted feature instances of images. The goal of ZSL is to learn a model f to classify unseen samples during the test phase, which can be formulated as $f : x \rightarrow \mathcal{Y}^u$. GZSL is a more realistic and challenging problem that requires f to handle both seen and unseen samples: $f : x \rightarrow \mathcal{Y}^{all}$.

3.2. Task-Shared and Task-Specific Representations

To begin our PS-GZSL, we first provide definitions for three visual representations that are concerning discriminative and semantic-relevant concepts.

Discriminative and Semantic-relevant Representations. Firstly, we define task-shared discriminative and semantic-relevant representations h_{ds} to encode the discriminative features of images that are related to corresponding semantic descriptors. These visual features are used for the both discrimination task and the visual-semantic alignment task during the training phase.

Discriminative but Non-semantic Representations. Secondly, discriminative but non-semantic features are encoded in discrimination task-specific representations, denoted as h_d. These features are important for discrimination, but they may not contribute to the visual-semantic alignment task since not represented in the semantic descriptors.

Non-Discriminative but Semantic-relevant Representations. Finally, non-discriminative but semantic-relevant features are encoded in visual-semantic alignment task-specific representations, denoted as h_s. These features are not discriminative in seen classes but may be critical for recognizing unseen classes. Thus, these features only contribute to the visual-semantic alignment task during training.

3.3. Representation Learning

As shown in Figure 2, Our encoder module consists of three parallel Mixture-of-experts (MoE) modules ($E = [E_{ds}, E_d, E_s]$), which explicitly factorize a visual feature x into three latent representations: h_{ds}, h_d, and h_s, i.e., $h_{ds} = E_{ds}(x)$, $h_d = E_d(x)$ and $h_s = E_s(x)$.

3.3.1. Mixture-of-Experts

PS-GZSL adopts a gated MoE module to replace simple Multi-Layer Perceptrons (MLPs) in order to obtain more expressive representations, MoE is a neural network architecture that comprises several experts, each of which specializes in a specific part of the input space. The output of the network is then computed as a weighted combination of the outputs of the experts by a gating network, as shown in Figure 3.

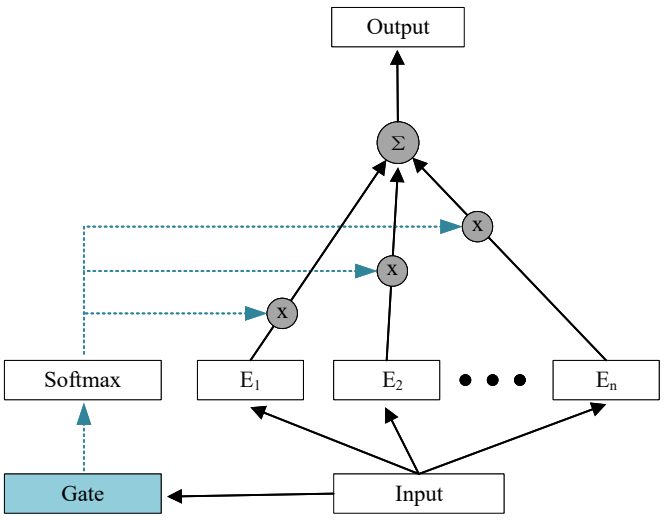


Figure 3. The architecture of MoE.

Given a visual feature as input, the MoE module can be formulated as:

$$E(x) = \sum_{i}^{n} g(x)_i e_i(x), \tag{1}$$

where, the gate network g combines the results of n expert networks, where $\sum_{i=1}^{n} g(x)_i = 1$ and $g(x)_i$ represents the ith logit of the output, indicating the weight assigned to expert e_i.

We denote the aforementioned three MoE modules as E_{ds}, E_d and E_s for the task-shared representation h_{ds} and two task-specific representation h_d and h_s, respectively. It's worth noting that we've incorporated the dropout technique in the gate network, which randomly discards some outputs of the experts. This technique helps prevent overfitting and also ensures that the representations (h_{ds}, h_d, and h_s) remain informative for subsequent sub-tasks.

3.3.2. Instance Contrastive Discrimination Task

According to the definition above, both h_{ds} and h_d are expected to capture the discriminative features. For convenience, we denote $w = h_{ds} \oplus h_d = E_{ds}(x) \oplus E_d(x)$. To compare the similarities and differences of visual representations w, an instance contrastive discrimination task is proposed, which assigns samples to different categories according to the comparison results. Specifically, PS-GZSL takes Supervised Contrastive Learning (SupCon) [19] loss as the objective function in this task since SupCon shows better generalization performance and stronger robustness in discriminative representation learning compared with other metric learning loss.

We follow the strategy proposed in [19] where the representation w is further propagated through a projection network P (as shown in Figure 4) to obtain a new representation denoted as $z = P(w)$. For every w_i encoded from a visual feature x_i, the SupCon loss of w_i is as follows:

$$\ell(z_i) = -\log\left\{\frac{1}{P(i)} \sum_{p \in P_i} \frac{\exp\left(z_i^\top z_p / \tau_e\right)}{\sum_{k \in K(i)} \exp\left(z_i^\top z_k / \tau_e\right)}\right\}, \tag{2}$$

where, $\tau_e > 0$ denotes the temperature parameter for stable training. $P(i) \equiv \{p \in K(i) : \boldsymbol{y}_p = \boldsymbol{y}_i\}$ represents the indices of all positives in the mini-batch that are distinct from i, and $|P(i)|$ is its cardinality.

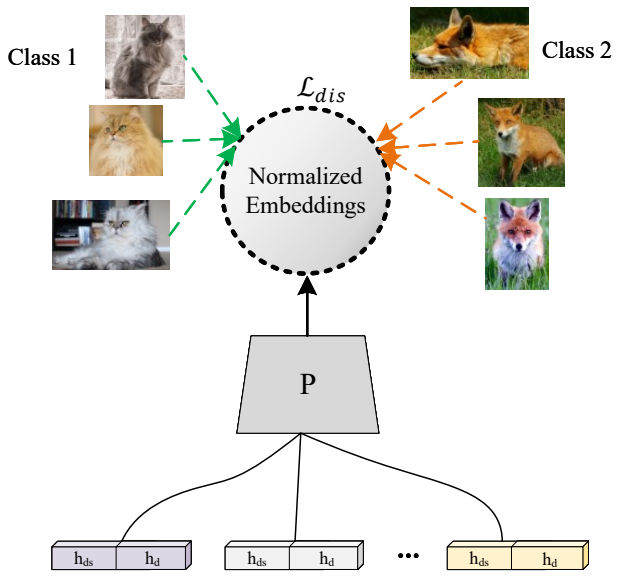


Figure 4. Illustration of Instance Contrastive Discrimination.

To simultaneously learn the MoE modules E_{ds}, E_d, and the projection network P, the loss function for this discrimination task is calculated as the sum of instance-level SupCon loss within a batch of samples I.

$$\mathcal{L}_{dis}(E_d, E_{ds}, P) = \sum_{i \in I} \ell(z_i). \tag{3}$$

Such a contrastive learning encourages E_{ds} and E_d to capture the strong inter-class discriminative features, and intra-class structure shared in the latent space, making both h_{ds} and h_d more discriminative and more transferable. Furthermore, we demonstrate the superiority of SupCon loss over softmax loss in ablation experiments.

3.3.3. Relation-Based Visual-Semantic Alignment Task

In the same way, both h_{ds} and h_s are devised to capture semantic-relevant information that corresponds to the annotated semantic descriptors A. For convenience, we denote $v = h_{ds} \oplus h_s = E_{ds}(x) \oplus E_s(x)$. In order to learn semantic-relevant representations v without directly mapping visual features into the semantic space, we adopt a Relation network in [20] as a visual-semantic alignment task. The goal is to maximize the similarity score (SS) between v and the corresponding semantic descriptor a through a deeper end-to-end architecture, which includes a learned nonlinear metric in the form of our alignment task. Thus, the objective of this task is to accurately measure the similarity score between pairs of v and a via a neural network. The similarity score SS of the matched pairs is set to 1, while mismatched pairs are assigned 0, which can be formulated as:

$$SS(v_t, a_c) = \begin{cases} 0, & y_t \neq y_c \\ 1, & y_t = y_c \end{cases}, \tag{4}$$

where t and c refer to the t-th visual sample's semantic-relevant representation and c-th class-level semantic descriptor from the seen classes, y_t and y_c denote the ground truth label of v_t and a_c.

In [20], they utilize mean square error(MSE) as a loss function while ignoring the class-imbalance problem in zero-shot learning. Moreover, as SupCon requires a large batch size, "Softmax + Cross Entropy" is a more efficient alternative than MSE in this scenario (as shown in Figure 5).

Denote the relation module as R. We can calculate the loss function of this task as:

$$\mathcal{L}_{align}(E_s, E_{ds}, R) = \sum_{i \in I} -\log \frac{\exp\left(R(v_i, a^+)/\tau_s\right)}{\sum_{s=1}^{S} \exp\left(R(v_i, a_s)/\tau_s\right)}, \tag{5}$$

where, S denotes the number of seen classes, and $\tau_s > 0$ denotes the scaling factor to stable the softmax activation for robust performance.

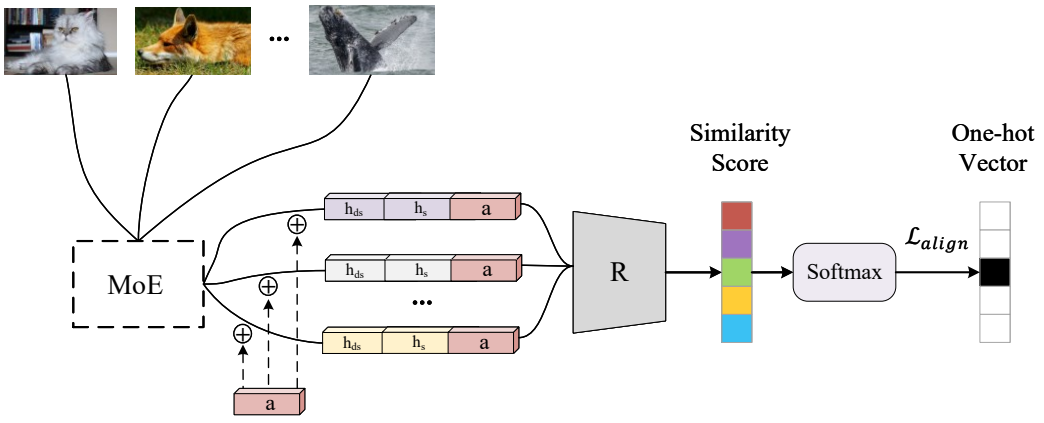


Figure 5. Illustration of Relation-Based Visual-Semantic Alignment.

3.4. Feature Generation with Latent Feedback

In order to alleviate the phenomenon that encoded representations are biased towards seen classes in GZSL, we integrate the proposed representation learning method on top of a conditional GAN (cGAN) [21]. Specifically, we adopt a conditional generator network G to generate virtual unseen features $\tilde{x} = G(a, \epsilon)$, here $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ represent a Gaussian noise. In the meanwhile, we train a discriminator D to distinguish between a real pair (x, a) and a generated pair $(\tilde{x}, a)$. The generator G and the discriminator D are jointly trained by minimizing the adversarial objective given as:

$$V(G, D) = \mathbb{E}_{p(x,a)}[\log D(x, a)] + \mathbb{E}_{p_G(\tilde{x},a)}[\log(1 - D(\tilde{x}, a))], \tag{6}$$

where $p_(x, a)$ and $p_G(\tilde{x}, a)$ represent the joint distribution of real/synthetic visual-semantic pairs, respectively.

However, the objective stated above does not guarantee that the generated features are discriminative or semantic-relevant. Drawing on the feedback mechanism in [13,21,39], we aim to improve the quality of generated features by passing them through the aforementioned multi-task network. Therefore, Equation (6) can be reformulated as:

$$\begin{aligned} V(G, D) =& \mathbb{E}_{p(x,a)}[\log D(x, a)] + \mathbb{E}_{p_G(\tilde{x},a)}[\log(1 - D(\tilde{x}, a))] \\ &+ \mathbb{E}_{p_G(\tilde{x},a)}[\delta_1 \mathcal{L}_{align} + \delta_2 \mathcal{L}_{dis}], \end{aligned} \tag{7}$$

3.5. Training and Inference

As a summary, the overall loss of our proposed method is formulated as:

$$\mathcal{L}_{total} = V(G, D) + \mathcal{L}_{dis}(E_d, E_{sh}, P) + \mathcal{L}_{align}(E_s, E_{sh}, R). \tag{8}$$

Given visual features and corresponding semantic descriptors from seen classes, PS-GZSL solves GZSL in four steps:

1. Training feature generation and representation learning models based on Equation (8).
2. These learned models are then used to synthesize and extract unseen class representations $\tilde{c}$.
3. Using real visual samples x from seen classes for training the partially-shared representation learning part and synthesized visual samples $\tilde{x}$ for tuning generator.
4. The final generalized zero-shot classifier is a single layer linear softmax classifier, learned on $\tilde{c}$ and c (extracted from real seen x and synthesized samples $\tilde{x}$), as depicted in Figure 6.

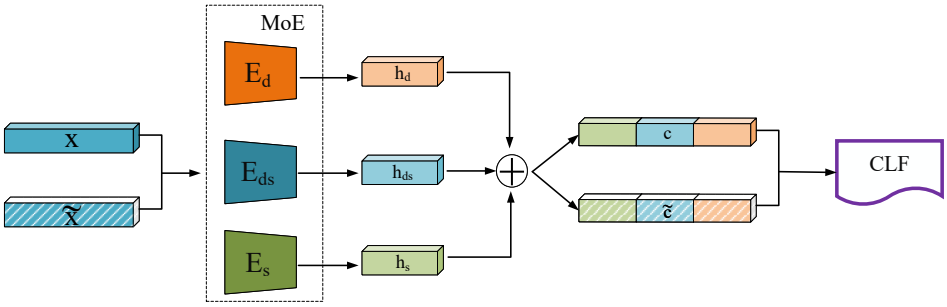


Figure 6. Using concatenated task-shared and task-specific representation for training classifier.

4. Experiments

4.1. Datasets

We perform our PS-GZSL on five widely used benchmark datasets for GZSL, including Animals with Attributes 1&2 (AWA1 [3] & AWA2 [1]), Caltech-UCSD Birds-200-2011 (CUB) [40], Oxford Flowers (FLO) [41]), and SUN Attribute (SUN) [42]. For visual features, we follow the standard GZSL practice of using ResNet101 [43] pre-trained on ImageNet-1k [44] without fine-tuning, resulting in 2048-dimensional features for each image. The semantic descriptors used for AWA1, AWA2, and SUN are their respective class-level attributes. For CUB and FLO, the semantic descriptors are generated from 10 textual descriptions by character-based CNN-RNN [45]. In addition, we employ the Proposed Split(PS) in [1] to split seen and unseen classes on each dataset. The statistics of the datasets and GZSL split settings are illustrated in Table 2.

Table 2. Statistics of the AWA1&2, CUB, and FLO, SUN datasets.

Dataset	AWA1	AWA2	CUB	FLO	SUN
#Seen Classes	40	40	150	82	645
#Unseen Classes	10	10	50	20	72
#Samples	30,475	37,322	11,788	8189	14,340
#Semantic Descriptors [1]	85	85	1024	1024	102
#Training Samples	19,832	23,527	7057	5394	10,320
#Test Seen Samples	4958	5882	1764	1640	2580
#Test Unseen Samples	5685	7913	2967	1155	1440

[1] #Semantic Descriptors indicate the dimensions of semantic descriptors per class.

4.2. Metrics

To assess the model performance in GZSL setting, we use the harmonic mean of per-class Top-1 accuracy on seen classes and unseen classes, formulated as $H = 2 \times S \times U/(S + U)$, where S and U represent seen accuracy and unseen accuracy, respectively. In addition, we adopt U as the evaluation metric for ZSL.

4.3. Implementation Details

In our PS-GZSL, all networks are implemented with Multi-Layer Perceptrons(MLPs). The architecture of the discriminator and generator of the feature generation architectures consist of single-layer MLPs with a 4096-unit hidden layer activated by LeakyReLU. In representation learning, each MoE module contains three experts and corresponds to a gate network. The dimension of task-specific representation (h_d & h_s) and task-shared representation (h_{sh}) are set to 1024 in all of the five datasets. For the projection network P, we set the size of the projection's output z to 256 for AWA2, FLO, and SUN and 512 for AWA1 and CUB. The relation network R contains two FC+ReLU layers, and we utilize 2048 hidden units for AWA1, AWA2, and CUB and 1024 units for FLO and SUN. The difference among datasets has motivated us to perform numerous experiments aimed at determining the optimal number of synthesized unseen visual instances in each dataset. Once PS-GZSL is trained, we use a fixed 400 per unseen class for CUB, 2400 for AWA1&2, 600 for FLO, and 100 for SUN. The weighting coefficients in Equation (7) are set to $\sigma_1 = 0.001$ and $\sigma_2 = 0.001$, and the value of temperature in Equations (2) and (5) are set to $\tau_e = 0.1$ and $\tau_s = 0.1$. We optimize the overall loss function (Equation (8)) with the Adam optimizer, using $\beta_1 = 0.5$, $\beta_2 = 0.999$. The mini-batch size is set to 512 for AWA1, AWA2, CUB, and SUN, and 3072 for FLO in our method. All experiments are implemented with PyTorch, and trained on a single NVIDIA RTX 2080Ti GPU.

4.4. Comparison with State-of-the-Arts

Recently, some methods have introduced transductive zero-shot learning on target datasets, where they use unlabeled unseen samples for training models, leading to sig-

nificant performance increases. However, it is costly and even unrealistic in real-world zero-shot scenarios. Thus, we only present results under the inductive setting.

Our PS-GZSL is compared with other GZSL methods on five widely used datasets without fine-tuning the pre-trained backbone. Results of our method in GZSL are given in Table 3, which indicates that PS-GZSL is compatible with the state-of-the-art. Specifically, PS-GZSL attains the best harmonic mean **H** on four datasets, i.e., 70.6 on AWA1, 71.8 on AWA2, 67.4 on CUB, and 43.3 on SUN. Notably, on CUB, PS-GZSL is the first one that attains a performance > 70.0 on unseen accuracy, which is even higher than the seen accuracy. This is because PS-GZSL retains more information in the learned representations to enhance GZSL classification during testing. As a result, representations for seen classes contain some redundancy, which adversely affects their classification accuracy. On FLO, PS-GZSL achieves the second-best harmonic mean **H** with 73.8, only lower than FREE [14]. However, PS-GZSL outperforms FREE by a considerable margin on the other four datasets. These results show that PS-GZSL can acquire classification knowledge transferable to unseen classes by utilizing the partially-shared mechanism and MoE, thereby learning more transferable representations from the seen classes. Specifically, by explicitly preserving these task-specific representations, the three MoE modules can effectively reduce the loss of information caused by the conflict between discrimination and visual-semantic alignment, thus enabling the preservation of more useful features for the testing phase.

Table 3. Comparisons with the State-Of-The-Art GZSL Methods. The best results and the second-best results are respectively marked in red and blue.

Methods	AWA1			AWA2			CUB			FLO			SUN		
	U	**S**	**H**	**U**	**S**	**H**	**U**	**S**	**H**	**U**	**S**	**H**	**U**	**S**	**H**
DeViSE [22]	13.4	68.7	22.4	17.1	74.7	27.8	23.8	53.0	32.8	9.9	44.2	16.2	16.9	27.4	20.9
TCN [46]	49.4	**76.5**	60.0	61.2	65.8	63.4	52.6	52.0	52.3	-	-	-	31.2	37.3	34.0
DVBE [47]	-	-	-	63.6	70.8	67.0	53.2	60.2	56.5	-	-	-	45.0	37.2	40.7
f-CLSWGAN [21]	57.9	64.0	60.2	-	-	-	43.7	57.7	49.7	59.0	73.8	65.6	42.6	36.6	39.4
CADA-VAE [9]	57.3	72.8	64.1	55.8	75.0	63.9	51.6	53.5	52.4	-	-	-	47.2	35.7	40.6
SP-AEN [8]	-	-	-	23.3	**90.9**	37.1	34.7	**70.6**	46.6	-	-	-	24.9	38.6	30.3
LisGAN [28]	52.6	76.3	62.3	-	-	-	46.5	57.9	51.6	57.7	83.8	68.3	42.9	37.8	40.2
cycle-CLSWGAN [30]	56.9	64.0	60.2	-	-	-	45.7	61.0	52.3	59.2	72.5	65.1	**49.4**	33.6	40.0
DLFZRL [11]	-	-	61.2	-	-	60.9	-	-	51.9	-	-	-	-	-	42.5
cvcZSL [48]	62.7	**77.0**	**69.1**	56.4	**81.4**	66.7	47.4	47.6	47.5	-	-	-	36.3	42.8	39.3
f-VAEGAN-D2 [29]	57.9	61.4	59.6	-	-	-	43.7	57.7	49.7	59.0	73.8	65.6	42.6	36.6	39.4
LsrGAN [31]	54.6	74.6	63.0	-	-	-	48.1	59.1	53.0	-	-	-	44.8	37.7	40.9
TF-VAEGAN [39]	-	-	-	59.8	75.1	66.6	52.8	64.7	58.1	62.5	**84.1**	71.7	45.6	**40.7**	43.0
DR-GZSL [7]	60.7	72.9	66.2	56.9	80.2	66.6	51.1	58.2	54.4	-	-	-	36.6	**47.6**	41.4
SDGZSL [10]	-	-	-	**64.6**	73.6	68.8	59.9	66.4	63.0	62.2	79.3	69.8	48.2	36.1	41.3
CE-GZSL [13]	**65.3**	73.4	**69.1**	63.1	78.6	**70.0**	**63.9**	**66.8**	**65.3**	**69.0**	78.7	73.5	48.8	38.6	**43.1**
FREE [14]	62.9	69.4	66.0	60.4	75.4	67.1	55.7	59.9	57.7	**67.4**	**84.5**	**75.0**	47.4	37.2	41.7
Our PS-GZSL	**67.5**	74.1	**70.6**	**66.4**	78.1	**71.8**	**70.6**	64.5	**67.4**	66.8	82.5	**73.8**	**50.1**	38.1	**43.3**

Furthermore, we also report the performances of our PS-GZSL in the conventional ZSL scenario, as presented in Table 4. To provide a comprehensive comparison, we have selected both previous conventional ZSL methods and recent GZSL methods under the conventional zero-shot setting. PS-GZSL achieves the best performance on three datasets and the second-best on FLO and SUN. This shows its superiority over existing GZSL methods on unseen classes and its strong generalization ability. These results prove the effectiveness of our PS-GZSL in both GZSL and conventional ZSL.

Table 4. Results of conventional ZSL. The best and the second-best accuracy of unseen classes are respectively marked in red and blue.

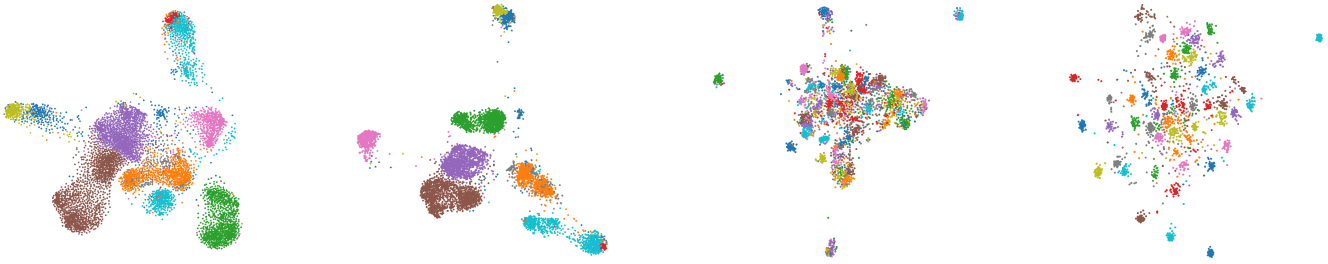
Methods	AWA1	AWA2	CUB	FLO	SUN
DEVISE [22]	54.2	59.7	52.0	45.9	56.5
SJE [23]	65.6	61.9	53.9	53.4	53.7
ALE [24]	59.9	62.5	54.9	48.5	58.1
ESZSL [25]	58.2	58.6	53.9	51.0	54.5
DCN [26]	65.2	-	56.2	-	61.8
CADA-VAE [9]	-	64.0	60.4	65.2	61.8
SP-AEN [8]	58.5	-	55.4	-	59.2
cycle-CLSWGAN [30]	66.3	-	58.4	70.1	60.0
DLFZRL [11]	**71.3**	70.3	61.8	-	61.3
TCN [46]	70.3	71.2	59.5	-	61.5
f-CLSWGAN [21]	68.2	-	57.3	67.2	60.8
f-VAEGAN-D2 [29]	-	71.1	61.0	67.7	**64.7**
TF-VAEGAN [39]	-	72.2	64.9	70.8	**66.0**
AGZSL [12]	-	**72.8**	76.0	-	63.3
SDGZSL [10]	-	72.1	75.5	**73.3**	62.4
CE-GZSL [13]	71.0	70.4	**77.5**	70.6	63.3
Ours PS-GZSL	**71.5**	**72.9**	**78.1**	**71.3**	**64.7**

4.5. Ablation Studies

Ablation studies were conducted to gain further insight into our PS-GZSL, evaluating the effects of different model architectures and representation components.

4.5.1. t-SNE Visualization

To further validate the transferability of our PS-GZSL, we visualize the task-shared representation h_{ds} and the multi-task joint representation $h_{ds} \oplus h_d \oplus h_s$ from unseen visual samples in Figure 7. We choose 10 unseen categories of test unseen set on AWA2 and 50 unseen categories of test unseen set on CUB. These data are sufficient in quantity and explicitly show the model's learned representation for the class comparison in unseen classes. Clearly, as we expected, the multi-task joint representation is more discriminative than the individual task-shared representation. However, we can still see discriminative patterns from h_{ds}, which is consistent with the assumption of previous methods based on learning the shared parts. This demonstrates that these task-shared representations may help classify between these categories, but the discriminative knowledge transfer from known to unknown categories is impaired due to the loss of task-specific information.



(**a**) (**b**) (**c**) (**d**)

Figure 7. The t-SNE visualization:. (**a**) h_{ds} of unseen classes on AWA2, (**b**) $h_{ds} \oplus h_d \oplus h_s$ of unseen classes on AWA2, (**c**) h_{ds} of unseen classes on CUB and (**d**) $h_{ds} \oplus h_d \oplus h_s$ of unseen classes on CUB.

4.5.2. Effectiveness of Task-Shared & Task-Specific Representations

In order to validate our key motivation for the partially-shared mechanism of PS-GZSL: In addition to task-shared *discriminative and semantic-relevant* representations, task-specific *only discriminative* representations and *only semantic-relevant* representations are both useful

in GZSL. We studied the performance of different combinations among h_{ds}, h_d and h_s. The results are presented in Figure 8, where we observe that using h_{ds} alone achieves comparable poor performance. However, when h_{ds} is concatenated with either h_d or h_s, the performance is improved, which demonstrates that both the h_d and h_s are helpful in GZSL. The best performance is achieved when we concatenate h_{ds}, h_d, and h_s together. This reveals that task joint representation $h_{ds} \oplus h_d \oplus h_s$ can capture complete correlation information among categories and their semantic descriptors, resulting in more informative and transferable representations for the test phase. Thus, both the task-shared and task-specific representations between discrimination and visual-semantic alignment are crucial to improve the classification performance in GZSL.

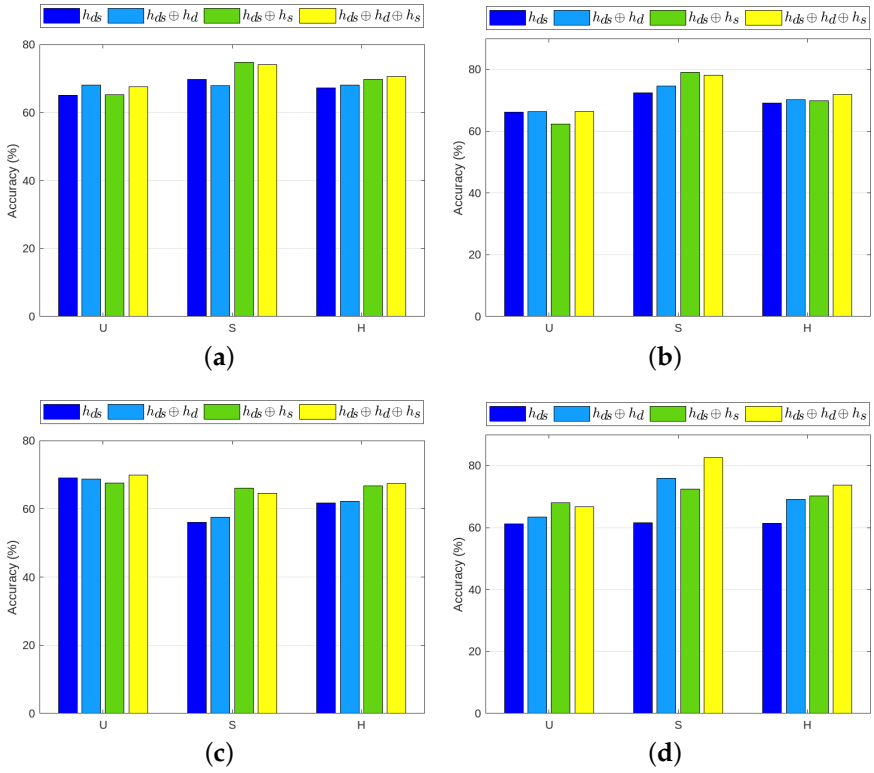


Figure 8. The effectiveness of various latent representations: (**a**) AWA1, (**b**) AWA2, (**c**) CUB and (**d**) FLO.

4.5.3. Analysis of Model Components

To assess the contributions of each component in PS-GZSL, different stripped-down architectures of we proposed methods were evaluated. The GZSL performance of each version on the AWA2 and CUB is represented in Table 5.

We observe that PS-GZSL outperforms PS-GZSL w/o MoE which validates that the MoE can improve the transferability of representation in GZSL. More importantly, we observe that PS-GZSL w/o MoE&PS outperforms PS-GZSL w/o MoE. This reveals the fact that simply splitting the visual encoder into three branches is not sufficient for learning the ideal transferable representations. Because any arbitrary mutually exclusive information decomposition can satisfy the regularizer, even if the h_{ds} encodes total information and h_d, h_s are non-informative for both tasks. This further demonstrates the superiority of our MoE module and expert dropout mechanism, which avoids the inexpressive issue among h_{ds}, h_d, and h_s. The above results indicate that our partially-shared mechanism and MoE module are mutually complementary in our method and prove that jointly preserving shared and specific representations between discriminative features and semantic features can preserve more complete and transferable information.

Table 5. Ablation study for different stripped-down architectures of PS-GZSL on the AWA2 and CUB dataset. PS is the partially-shared mechanism, $\mathcal{L}_{dis}$ is the adopted SupCon loss, $\mathcal{L}_{clf}$ is a classification loss of an auxiliary classifier for our discrimination task, and $\mathcal{L}_{mse}$ is the MSE version of our visual-semantic alignment task. The best and the second-best accuracy of unseen classes are respectively marked in red and blue.

Version	AWA2			CUB		
	U	**S**	**H**	**U**	**S**	**H**
PS-GZSL w/o MoE&PS	65.7	74.8	69.9	**71.5**	61.3	66.0
PS-GZSL w/o PS	**66.9**	74.8	70.7	67.0	**66.8**	66.9
PS-GZSL w/o MoE	61.4	**79.8**	69.4	68.4	63.1	65.6
PS-GZSL w/o $\mathcal{L}_{align}$ w/ $\mathcal{L}_{mse}$	66.0	75.5	70.5	66.9	66.2	66.5
PS-GZSL w/o $\mathcal{L}_{dis}$ w/ $\mathcal{L}_{clf}$	65.7	77.8	**71.2**	67.5	**66.8**	**67.2**
PS-GZSL	**66.4**	**78.1**	**71.8**	**70.1**	64.5	**67.4**

4.6. Hyper-Parameter Analysis

In our PS-GZSL approach, the hyperparameters that exert the greatest influence are the number of synthesized samples per class, the number of experts in each branch, and the dimensions of h_{ds}, h_d, and h_s.

Visualization of Different Number of Synthesized Samples. The number of synthesized samples per class was varied, as shown in Figure 9. The results show that the performance on all four datasets increased with an increasing number of synthesized examples. This demonstrated that the bias towards seen problems was relieved by the feature generation in our PS-GZSL. However, generating too many samples will impair the accuracy of seen classes (**S**) and eventually hamper the harmonic mean **H**. Therefore, selecting an appropriate value to achieve the balance between **S** and **U** is important.

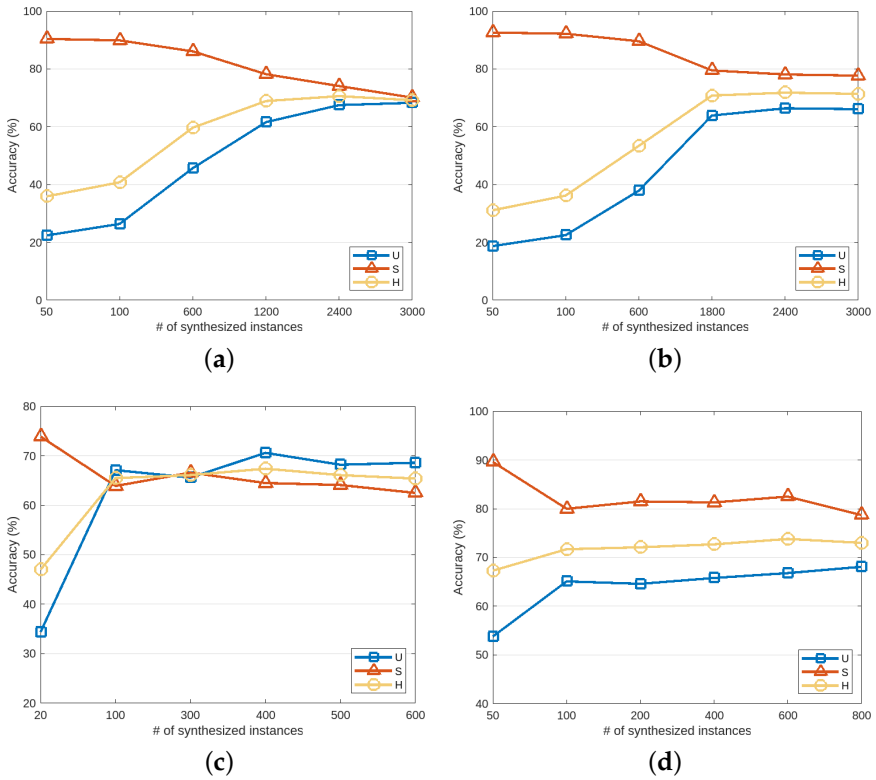
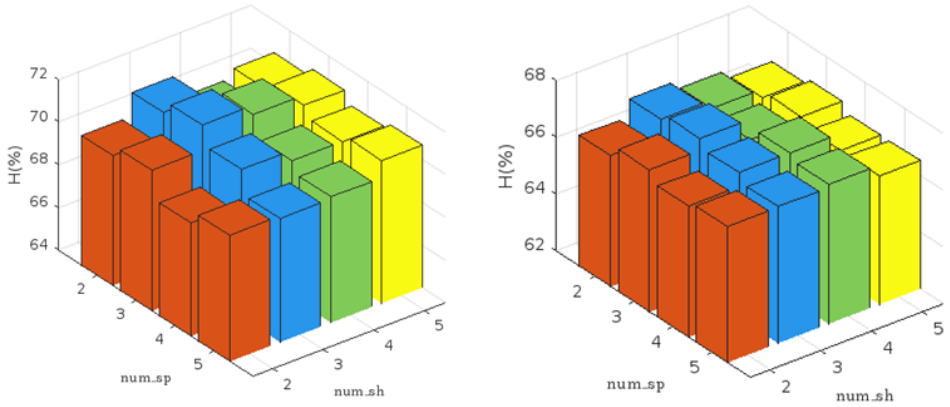


Figure 9. The influence of the number of synthesized visual instances in each unseen class. (**a**) AWA1, (**b**) AWA2, (**c**) CUB, and (**d**) FLO.

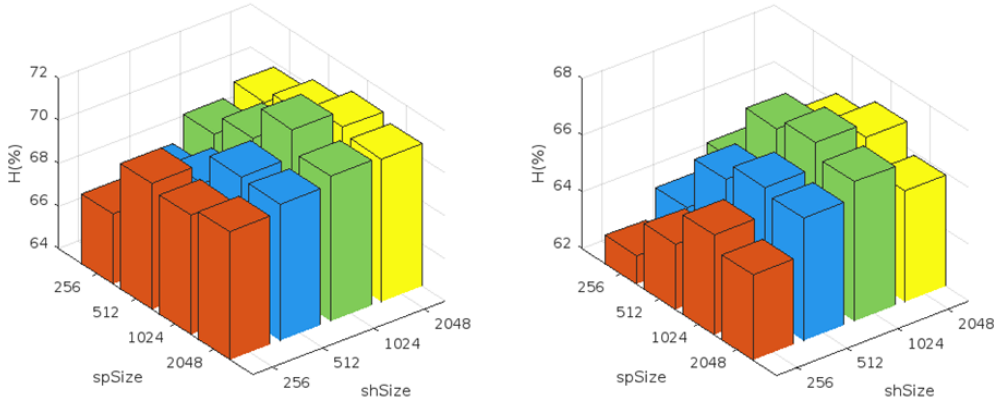
Visualization of Different Number of Experts. Since we use MoE modules for each branch, the architecture of the expert network is very important for our method. As shown in Figure 10, we study different numbers of experts for task-specific and task-shared, noted as **num_sp** and **num_sh**, respectively. As the numbers of task-specific experts and task-shared experts increase, the harmonic mean is boosted and then drops, which achieves the peek performance when num_sp = 3 and num_sh = 3. Thus, for convenience, both num_sp and num_sh are set to 3 in order to achieve a considerable performance in all of the remaining datasets.



(**a**) (**b**)

Figure 10. The effect of the number of task-specific and task-shared experts (denoted as **num_sp** and **num_sh**, respectively) : (**a**) AWA2 and (**b**) CUB.

Visualization of Different Representations Dimensions. Intuitively, the dimensions h_{ds}, h_d, and h_s will have a significant impact on the optimization of these two sub-tasks. This will ultimately affect the transferability and expressiveness of the concatenated final representations. To explore the sensitivity of our PS-GZSL to the dimensionality in the latent space. As shown in Figure 11, the harmonic mean accuracy of PS-GZSL for different latent dimensions on AWA2 and CUB, i.e., 256, 512, 1024, and 2048 for both task-specific and task-shared representations(denoted as **spSize** and **shSize**, respectively) are represented. As **spSize** and **shSize** are both set to 1024, PS-GZSL consistently performs better than all others on AWA2 and CUB. Therefore, both **spSize** and **shSize** are set to 1024 in all of the remaining datasets.



(**a**) (**b**)

Figure 11. The effect of the dimension of task-specific representation and task-shared representation (denoted as **spSize** and **shSize**, respectively). (**a**) AWA2 and (**b**) CUB.

5. Conclusions

In this paper, we propose a new way of learning the composite method by accounting for all the features based on multi-task representation learning. Specifically, the recent representation learning method in GZSL discards some specific information between two tasks (i.e., classification task and visual semantic alignment task). As explained in the introduction, this specific information can be either discriminative or semantic-relevant, depending on their contribution to the testing phase.

Further on, we believe that jointly preserving task-specific and task-shared features leads to a more complete and more transferable representation in GZSL. To support this claim, a novel representation learning method termed PS-GZSL is proposed. Unlike most existing methods, PS-GZSL explicitly factorizes visual features into one task-shared and two task-specific representations through the partially-shared mechanism between the discrimination and visual semantic alignment task. This flexibility enables PS-GZSL to preserve more complete knowledge. Furthermore, PS-GZSL carefully designs the mixture of experts and gate networks for learning informative representations for each branch. As evaluated in extensive experiments, the good transferability of PS-GZSL has been demonstrated.

As a starting point, this study shows the potential ability of the partially-shared mechanism in learning transferable representation in GZSL. There is still a large research space in this direction. First, the relative loss weight ratio of each sub-task is set to 1, but future work could investigate the use of adaptive weights to balance the two tasks during optimization. Second, ideally, the encoding information of task-shared and task-specific representations should be no redundancy. It is also important to devise a regularizer to accomplish this. In the future, we will investigate these potential directions.

Author Contributions: Conceptualization, G.W. and S.T.; methodology, G.W.; software, G.W.; validation, S.T.; formal analysis, G.W. and S.T.; investigation, G.W.; resources, G.W.; data curation, G.W.; writing—original draft preparation, G.W.; writing—review and editing, S.T.; visualization, G.W.; supervision, S.T.; project administration, G.W.; funding acquisition, S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available upon request from the first author.

Acknowledgments: We are grateful for resources from the High-Performance Computing Center of Central South University.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ZSL	Zero-Shot Learning
GZSL	Generalized Zero-Shot Learning
SupCon	Supervised Contrastive
MoE	Mixture-of-Experts
PS	Partially-Shared mechanism

References

1. Xian, Y.; Schiele, B.; Akata, Z. Zero-shot learning-the good, the bad and the ugly. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4582–4591.
2. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
3. Lampert, C.H.; Nickisch, H.; Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 951–958.

4. Palatucci, M.; Pomerleau, D.; Hinton, G.E.; Mitchell, T.M. Zero-shot learning with semantic output codes. *Adv. Neural Inf. Process. Syst.* **2009**, *22*, 1410–1418.
5. Chao, W.L.; Changpinyo, S.; Gong, B.; Sha, F. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part II 14*; Springer: Cham, Switzerland, 2016; pp. 52–68.
6. Saad, E.; Paprzycki, M.; Ganzha, M.; Bădică, A.; Bădică, C.; Fidanova, S.; Lirkov, I.; Ivanović, M. Generalized Zero-Shot Learning for Image Classification—Comparing Performance of Popular Approaches. *Information* **2022**, *13*, 561. [CrossRef]
7. Li, X.; Xu, Z.; Wei, K.; Deng, C. Generalized zero-shot learning via disentangled representation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 1966–1974.
8. Chen, L.; Zhang, H.; Xiao, J.; Liu, W.; Chang, S.F. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1043–1052.
9. Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; Akata, Z. Generalized zero-and few-shot learning via aligned variational autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 8247–8255.
10. Chen, Z.; Luo, Y.; Qiu, R.; Wang, S.; Huang, Z.; Li, J.; Zhang, Z. Semantics disentangling for generalized zero-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 8712–8720.
11. Tong, B.; Wang, C.; Klinkigt, M.; Kobayashi, Y.; Nonaka, Y. Hierarchical disentanglement of discriminative latent features for zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 11467–11476.
12. Chou, Y.Y.; Lin, H.T.; Liu, T.L. Adaptive and generative zero-shot learning. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
13. Han, Z.; Fu, Z.; Chen, S.; Yang, J. Contrastive embedding for generalized zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Conference, 19–25 June 2021; pp. 2371–2381.
14. Chen, S.; Wang, W.; Xia, B.; Peng, Q.; You, X.; Zheng, F.; Shao, L. Free: Feature refinement for generalized zero-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 122–131.
15. Bui, M.H.; Tran, T.; Tran, A.; Phung, D. Exploiting domain-specific features to enhance domain generalization. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 21189–21201.
16. Milbich, T.; Roth, K.; Bharadhwaj, H.; Sinha, S.; Bengio, Y.; Ommer, B.; Cohen, J.P. Diva: Diverse visual feature aggregation for deep metric learning. In *Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020, Proceedings, Part VIII 16*; Springer: Cham, Switzerland, 2020; pp. 590–607.
17. Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; Chi, E.H. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 1930–1939.
18. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive mixtures of local experts. *Neural Comput.* **1991**, *3*, 79–87. [CrossRef] [PubMed]
19. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
20. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; Hospedales, T.M. Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1199–1208.
21. Xian, Y.; Lorenz, T.; Schiele, B.; Akata, Z. Feature generating networks for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5542–5551.
22. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; Mikolov, T. Devise: A deep visual-semantic embedding model. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 2121–2129.
23. Akata, Z.; Reed, S.; Walter, D.; Lee, H.; Schiele, B. Evaluation of output embeddings for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2927–2936.
24. Akata, Z.; Perronnin, F.; Harchaoui, Z.; Schmid, C. Label-embedding for attribute-based classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 819–826.
25. Romera-Paredes, B.; Torr, P. An embarrassingly simple approach to zero-shot learning. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2152–2161.
26. Liu, S.; Long, M.; Wang, J.; Jordan, M.I. Generalized zero-shot learning with deep calibration network. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 2009–2019.
27. Yang, G.; Han, A.; Liu, X.; Liu, Y.; Wei, T.; Zhang, Z. Enhancing Semantic-Consistent Features and Transforming Discriminative Features for Generalized Zero-Shot Classifications. *Appl. Sci.* **2022**, *12*, 12642. [CrossRef]

28. Li, J.; Jing, M.; Lu, K.; Ding, Z.; Zhu, L.; Huang, Z. Leveraging the invariant side of generative zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7402–7411.
29. Xian, Y.; Sharma, S.; Schiele, B.; Akata, Z. f-vaegan-d2: A feature generating framework for any-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10275–10284.
30. Felix, R.; Reid, I.; Carneiro, G. Multi-modal cycle-consistent generalized zero-shot learning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 21–37.
31. Vyas, M.R.; Venkateswara, H.; Panchanathan, S. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In *Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020, Proceedings, Part XXX 16*; Springer: Cham, Switzerland, 2020; pp. 70–86.
32. Li, Z.; Zhang, D.; Wang, Y.; Lin, D.; Zhang, J. Generative Adversarial Networks for Zero-Shot Remote Sensing Scene Classification. *Appl. Sci.* **2022**, *12*, 3760. [CrossRef]
33. Sohn, K.; Lee, H.; Yan, X. Learning structured output representation using deep conditional generative models. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 3483–3491.
34. Verma, V.K.; Arora, G.; Mishra, A.; Rai, P. Generalized zero-shot learning via synthesized examples. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4281–4289.
35. Kim, J.; Shim, K.; Shim, B. Semantic feature extraction for generalized zero-shot learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 1166–1173.
36. Tang, H.; Liu, J.; Zhao, M.; Gong, X. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In Proceedings of the 14th ACM Conference on Recommender Systems, Virtual Event, 22–26 September 2020; pp. 269–278.
37. Park, H.; Yeo, J.; Wang, G.; Hwang, S.W. Soft representation learning for sparse transfer. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 1560–1568.
38. Xin, S.; Jiao, Y.; Long, C.; Wang, Y.; Wang, X.; Yang, S.; Liu, J.; Zhang, J. Prototype Feature Extraction for Multi-task Learning. In Proceedings of the ACM Web Conference 2022, Lyon France, 25–29 April 2022; pp. 2472–2481.
39. Narayan, S.; Gupta, A.; Khan, F.S.; Snoek, C.G.; Shao, L. Latent embedding feedback and discriminative features for zero-shot classification. In *Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020, Proceedings, Part XXII 16*; Springer: Cham, Switzerland, 2020; pp. 479–495.
40. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. The Caltech-Ucsd Birds-200-2011 Dataset. 2011. Available online: https://authors.library.caltech.edu/27452/ (accessed on 29 March 2023).
41. Nilsback, M.E.; Zisserman, A. Automated flower classification over a large number of classes. In Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, Bhubaneswar, India, 16–19 December 2008; pp. 722–729.
42. Patterson, G.; Hays, J. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2751–2758.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
45. Reed, S.; Akata, Z.; Lee, H.; Schiele, B. Learning deep representations of fine-grained visual descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 49–58.
46. Jiang, H.; Wang, R.; Shan, S.; Chen, X. Transferable contrastive network for generalized zero-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9765–9774.
47. Min, S.; Yao, H.; Xie, H.; Wang, C.; Zha, Z.J.; Zhang, Y. Domain-aware visual bias eliminating for generalized zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12664–12673.
48. Li, K.; Min, M.R.; Fu, Y. Rethinking zero-shot learning: A conditional visual classification perspective. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3583–3592.