Improving Generalized Zero-Shot Learning by Exploring the Diverse Semantics from External Class Names

Yapeng Li[1] Yong Luo[1,2] Zengmao Wang[1] Bo Du[1,2*]

[1] National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, Hubei Key Laboratory of Multimedia and Network Communication Engineering, School of Computer Science, Wuhan University, Wuhan, China.

[2] Hubei Luojia Laboratory, Wuhan, China.

https://github.com/li-yapeng/DESCN

Abstract: Generalized Zero-Shot Learning (GZSL) methods often assume that the unseen classes are similar to seen classes, and thus perform poor when unseen classes are dissimilar to seen classes. Although some existing GZSL approaches can alleviate this issue by leveraging additional semantic information from test unseen classes, their generalization ability to dissimilar unseen classes is still unsatisfactory. In this paper, we propose a simple yet effective GZSL framework by exploring diverse semantics from external class names (DSECN), which is simultaneously robust on the similar and dissimilar unseen classes. **这篇文章来自 CVPR2024，仅为练习 latex 的语法、操作使用，由于技术相似，本文并未复现文章所有内容，仅提取了一张图片、一个表格、一个算法、若干公式、部分参考文献，以及实现源码地址跳转、参考文献跳转功能等。通过本项目的实践基本体会了 latex 的强大，选择临摹本文的原因是本文展示的内容种类较为齐全，方便练习。在此向原作者表示深深的感谢。**

1 Introduction

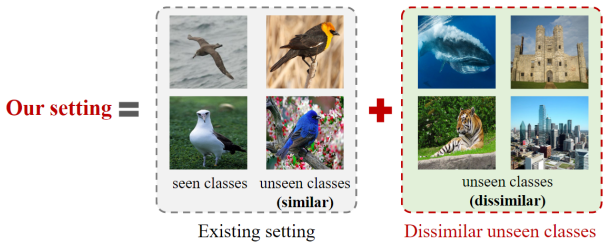
Due to the high cost of annotation and the complexity of real-world test scenarios, the presence of unseen classes is often inevitable.Unfortunately, traditional machine learning models are unable to handle samples from classes that have not been covered by the training data.

To tackle this challenge, zero-shot learning (ZSL) has been proposed to recognize new classes via transferring knowledge obtained from seen classes with the help of semantic information [[4], [5], [6], [7]]. In contrast, Generalized ZSL (GZSL) is a more challenging task which handle test samples from both seen and unseen classes .

The core idea of GZSL is to introduce auxiliary semantic information and establish the connection between semantic and visual space for the recognition of unseen classes . Since unseen samples are not available during training, existing GZSL models often misclassify unseen class samples into seen classes during test (known as the bias issue) . To alleviate this issue, several strategies have been introduced. A common strategy is to utilize a backbone pre-trained on the ImageNet-1K dataset to extract visual features, which is beneficial to improve the generalization ability of visual features. However,the rich semantics that implicitly contained in the backbone classification heads are simply ignored in these approaches, and thus their ability to exploit diverse semantics is limited. Generative-based GZSL methods alleviate the bias problem by introducing the semantics from test unseen classes, and some semantic augmentation approaches focus on

enhancing the semantics of each class by leveraging the textual documents , large language model , etc . However, these approaches share a common limitation that they heavily rely on the semantics and visual features from seen classes to build the relations between visual and semantic space. This makes these models perform poorly on dissimilar unseen classes (see Fig. 1b), as little information can be transferred from seen to dissimilar unseen classes. That is, these methods can only deal with the unrealistic setting that the unseen classes are similar with seen classes see existing setting of Fig. 1a). However, it is inevitable that unseen classes may be quite different from seen classes in the real-world applications, and hence it is desirable to enhance GZSL models to effectively handle dissimilar unseen classes.

To achieve this goal, we need to address three main challenges: (I) how to obtain the information that can be transferred to dissimilar unseen classes; (II) how to enable effective information transfer without labeled training data; and (III) how to reduce costs as much as possible while ensuring effectiveness. If the cost of the solution exceed the cost of collecting the data and retraining the model, then such a solution would be meaningless. Therefore, we propose a simple yet effective GZSL framework by exploring Diverse Semantics from External Class Names (DSECN), which is robust on both the similar and dissimilar unseen classes. Specifically, we introduce the diverse semantics from external (not test unseen classes) class names as the bridge to reduce the gap between the seen and unseen classes, which is beneficial for the recognition of unseen classes (hallenge I) . Then we utilize the classification head pre-trained on large-scale dataset, e.g., ImageNet-1K, to align the introduced semantics to the visual space (challenge II). Finally, the hierarchical taxonomy of WordNet for the classes in large-scale dataset is introduced to further improve the diversity of semantics from class names. Since the class name and pre-trained classification head are quite easy to collect, the cost of our method is quite low.



(a) Our Setting

Mode	CN [8]	TZero [3]	ZLA [1]	DGZ [2]	Ours (Δ)
Similar	42.50	47.99	42.65	43.16	**48.45** (+0.46)
Dissimilar	5.09	2.47	8.52	12.30	**47.44** (+35.14)

(b) Results

图 1: Figure 1. Setting illustration and results: (a) Comparison with the existing setting, our setting takes into account the scenario where the unseen classes are dissimilar to seen classes; (b) Performance of existing GZSL methods on similar and dissimilar unseen classes on the CUB dataset. The similar unseen classes were obtained from the same CUB dataset, while the dissimilar unseen classes come from AWA2 and SUN dataset.

To summarize, the main contributions of this paper are:

- To the best of our knowledge, we are the first that explicitly study the realistic GZSL setting that both similar and dissimilar unseen classes exist.
- We propose a GZSL method that explores the diverse semantics from external class names (DSECN), which is simultaneously robust on the similar and dissimilar unseen classes.
- We show that the proposed idea can be easily integrated into other GZSL approaches, such as generative-based ones, and improve their robustness for dissimilar unseen classes.

We conduct extensive experiments on diverse real-world datasets. The results show that in the practical setting including both similar and dissimilar unseen classes, the harmonic mean accuracies of our method significantly outperform all counter-

parts. Besides, our model can be trained within one minute on all three datasets.

2 Proposed Method

Problem Formulation. In ZSL and GZSL, we define two sets of classes: seen classes in Y^s and unseen classes in Y^u. The seen classes Y^s and the unseen classes Y^u are disjoint, i.e., $Y^s \cap Y^u = \emptyset$ and $Y^s \cup Y^u = \mathcal{Y}$. The unseen classes may be similar or dissimilar to seen classes in practice. Hence, in this work, the unseen classes Y^u contain similar unseen classes Y^u_s and dissimilar unseen classes Y^u_d, i.e., $Y^u = Y^u_s \cup Y^u_d$.The seen data are expressed as $D^s = \{(x^s_i, y^s_i)\}$, where $x^s_i \in \mathcal{X}$ indicates the i-th sample features extracted by the pretrained backbone network, e.g., ResNet101, and $y^s_i \in Y^s$ is its class label. The D^s is split into a training set D^s_{tr} and a testing set D^s_{te}. On the other hand, the unseen data are denoted as $D^u = \{(x^u_i, y^u_i)\}$, where $x^u_i \in \mathcal{X}^u$ and $y^u_i \in Y^u$ are the sample features of unseen classes and the corresponding ground-truth label for evaluation, respectively.The goal of ZSL is to learn a classifier for classifying test samples from unseen classes, i.e., $Xf_{ZSL}Y^u$. In contrast to ZSL, the goal of GZSL is to learn a classifier for classifying test samples from both seen and unseen classes, i.e., $Xf_{GZSL}Y^s \cup Y^u$.In ZSL and GZSL, the auxiliary semantic information $\mathcal{A}$ is obtained by transforming the class labels $\mathcal{Y}$ using human-annotated attributes [3] or language models [14, 15].

2.1 Visual Flow

The visual flow is designed to classify visual objects from both seen and unseen classes by transferring knowledge from the seen classes to the unseen ones with the help of semantic information. The visual flow contains two parts: semantic-to-visual subnetwork (S2V) and visual classifier. The S2V subnetwork is a learnable multilayer perceptron (MLP), and is used to link the semantic and visual representation. This enables the model to transfer the knowledge from seen classes to the unseen classes through semantic information. The visual classifier uses the relationship between the visual sample and all categories to obtain the classification result of the sample. Because the visual flow requires paired visual samples and category labels, in the training phase, the visual flow can only be trained on the paired visual feature and label dataset D^s_{tr}. Specifically, the S2V sub-network takes the seen semantic feature A^s as input to generate the class-level visual prototype of seen classes V^s. The process to generate the class-level visual prototype of seen classes V^s can be expressed by

$$A^s \in R^{C^s \times d^a} S2V V^s \in R^{C^s \times d^v}, \tag{1}$$

where C^s, d^a, d^v denote the number of seen classes, the dimensionality of semantic representation and the dimensionality of visual representation, respectively. A^s signifies the semantic representations of seen classes and is extracted by language models.

Next, the visual classifier (VC) computes the scaled cosine similarity between the visual feature of seen classes $X^s_{tr} \in R^{N^s_{tr} \times d^v}$ and the class-level visual feature of seen classes $V^s \in R^{C^s \times d^v}$ as logits $l^s \in R^{N^s_{tr} \times C^s}$, where N^s_{tr} is the number of seen class samples in the training set. Formally, the logits l^s can be obtained as follows:

$$l^s = \gamma^2 \cdot \frac{X^s_{tr} V^{s\top}}{\|X^s_{tr}\| \, \|V^s\|}, \tag{2}$$

where γ is the scale factor.

Finally, we adopt the Cross-Entropy (CE) loss to update the visual flow:

$$\mathcal{L}_V = \mathcal{L}_{CE}(\sigma(l^s), Y^s), \tag{3}$$

where σ denotes the *softmax* function.

2.2 Training and Inference

The proposed *DSECN*, which contains three components, is trained in an end-to-end manner, and the

total loss function is given as follows:

$$\mathcal{L}_{total} = \mathcal{L}_V + \lambda_{EB}\mathcal{L}_{EB} + \lambda_{IIA}\mathcal{L}_{IIA}, \quad (4)$$

where λ_{EB} and λ_{IIA} are trade-off hyper parameters. The training pseudo-code is presented in Algorithm 1. It is noteworthy that we only update the parameters of *S2V*.

In the inference phase, given a visual feature x and the semantic feature A of all classes, we apply the visual flow to obtain the final class prediction $\hat{y}$. Specifically, the semantic features $A = A^s \cup A^u$ of all classes are first fed into the *S2V* sub-network to generate the class-level visual prototype, i.e., $V = S2V(A)$. Then the final class prediction can be obtained according to:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \left(softmax \left(\gamma^2 \cdot \frac{xV^\top}{\|x\| \, \|V\|} \right) \right), \quad (5)$$

where $\mathcal{Y} = Y^s \cup Y^u$ is the union of seen classes Y^s and unseen classes Y^u.

2.3 Integrating into Existing GZSL Methods

The proposed DSECN can be easily integrated into other mainstream GZSL approaches to improve their robustness for dissimilar unseen classes. The GZSL method essentially models the relationship between visual and semantic features, so that semantic features can be used as a bridge to accurately classify visual features of unseen classes. DSECN can effectively build such relationship between "diverse" semantic and visual features at low cost, and significantly improving the performance of existing GZSL approaches. Taking generative-based methods as an example, the generator can generate visual features of the introduced diverse semantic from external class names. Then the pretrained classification head is used to align the generated visual features to visual space. This enables the generator to adapt to more classes and generate more accurate.

Algorithm 1: Training Process of DSECN

Input: Training seen data $\{X^s_{tr}, CN^s, Y^s_{tr}\}$, external base class names CN^{EB} and hierarchy augmented class names CN^{HA}
Output: The final S2V model

// Generate Semantics and Labels
$A^s, A^{EB}, A^{HA} = LM(CN^s, CN^{EB}, CN^{HA})$
$Y^{EB} \leftarrow \{CN^{EB}\}$ in Eq. (7)
$Y^{HA} \leftarrow \{CN^{HA}\}$ in Eq. (11)

for $e = 1, 2, ..., E$ **do**

// Visual Flow
$V^s = S2V(A^s)$
$l^s \leftarrow \{V^s, X^s_{tr}\}$ in Eq. (2)
$\mathcal{L}_V \leftarrow \{l^s, Y^s_{tr}\}$ in Eq. (3)

// Diverse Semantic Enhancement
$V^{EB} = S2V(A^{EB})$
$l^{EB} = SC(V^{EB})$
$\mathcal{L}_{EB} \leftarrow \{l^{EB}, Y^{EB}\}$ in Eq. (8)

// Hierarchy Taxonomy Enhancement
$V^{HA} = S2V(A^{HA})$
$l^{HA} = SC(V^{HA})$
$\mathcal{L}_{HA} \leftarrow \{l^{HA}, Y^{HA}\}$ in Eq. (12)

// Compute Total Loss
$\mathcal{L}_{total} = \mathcal{L}_V + \mathcal{L}_{EB} + \mathcal{L}_{HA}$

// Update Parameters
Update the parameters of S2V using Adam.
end

Finally, the refined visual features of unseen classes are used to train the GZSL classifier, thus improving the model performance. More details on the integration of our idea with the generative-based methods and other types of GZSL approaches can be found in the supplementary material A.

3 Conclusion

In this paper, we introduce and investigate the practical GZSL setting, where unseen classes can be either similar and dissimilar to seen classes. We empirically show that existing GZSL methods are difficult in identifying dissimilar unseen classes, and propose a simple yet effective method, which exploits diverse semantics from external class names (DSECN), and is simultaneously robust for both similar and dissimilar unseen classes. From the results, we mainly conclude that: 1) the semantics contained in the external unseen class names are quite helpful to improve the generalization ability of GZSL approach; 2) It is critical to align the augmented semantics to their corresponding visual features. In the future, we intend to incorporate the proposed idea into more GZSL approaches and improve their performance.

Potential Impacts. The existing GZSL methods perform poorly when unseen classes are dissimilar to seen classes, which hinders the practical application of the existing GZSL methods. The paper will help attract researchers' attention to dissimilar unseen classes and help the GZSL field develop in a more practical direction. In addition, the proposed DSECN is simultaneously robust on similar and dissimilar unseen classes, and can easily be integrated into other GZSL methods to improve their robustness for dissimilar unseen classes. This capability is beneficial for improving the practicality of the GZSL model. From an evaluation perspective, existing methods assess using visible and invisible classes from the same dataset. This leads to excessive focus on similar unseen classes during evaluation, thereby overestimating the generalizability of existing GZSL methods. In contrast, the cross-dataset evaluation protocol proposed in the paper can more comprehensively reflect the performance of existing GZSL methods on similar, dissimilar and practical settings.

Acknowledgement. This work was supported by the National Key Research and Development Program of China 2023YFC2705700, National Natural Science Foundation of China under Grants 62225113 and 62276195, and Special Fund of Hubei Luojia Laboratory under Grant 220100014.

References

[1] Dubing Chen, Yuming Shen, Haofeng Zhang, and Philip H.S. Torr. *Zero-shot logit adjustment.* In IJCAI, pages 813–819, 2022. Main Track.

[2] Dubing Chen, Yuming Shen, Haofeng Zhang, and Philip H.S. Torr. *Deconstructed generation-based zero-shot model.* In AAAI, pages 295–303, 2023.

[3] Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You. *TransZero: Attribute-guided transformer for zero-shot learning.* In AAAI, pages 330–338, 2022.

[4] Shay Deutsch, Soheil Kolouri, Kyungnam Kim, Yuri Owechko, and Stefano Soatto. *Zero shot learning via multiscale manifold regularization.* In CVPR, 2017.

[5] Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. *DeViSE: A deep visual-semantic embedding model.* In NeurIPS, 2013.

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

[7] Ziming Zhang and Venkatesh Saligrama. *Zero-shot learning via joint latent similarity embedding.* In CVPR, pages 6034–6042, 2016.

[8] Ivan Skorokhodov and Mohamed Elhoseiny. *Class normalization for (continual)? generalized zero-shot learning.* In ICLR, 2021.

[9] Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. *A survey of zero-shot learning: Settings, methods, and applications.* ACM Transactions on Intelligent Systems and Technology, 10(2):1–37, 2019.

[10] Jiamin Wu, Tianzhu Zhang, Zheng-Jun Zha, Jiebo Luo, Yongdong Zhang, and Feng Wu. *Self-supervised domain-aware generative network for generalized zero-shot learning.* In CVPR, pages 12767–12776, 2020.

[11] 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.

[12] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. *Zero-shot learning —a comprehensive evaluation of the good, the bad and the ugly.* IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 41(9):2251–2265, 2018.

[13] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. *Feature generating networks for zero-shot learning.* In CVPR, 2018.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. *Learning transferable visual models from natural language supervision.* In ICML, pages 8748–8763. PMLR, 2021.

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781, 2013.