Distilling Knowledge from Multiple Foundation Models for Zero-shot Image classification

Siqi Yin * and Lifan Jiang

Posted Date: 17 June 2024

doi: 10.20944/preprints202406.1153.v1

Keywords: Zero-shot learning; Multi-method integration; Image classification

 doi:10.20944/preprints202406.1153.v1

Article

Distilling Knowledge from Multiple Foundation Models for Zero-Shot Image Classification

Siqi Yin [1,*] and Lifan Jiang [2]

Shandong University of Science and Technology
* Correspondence: siqiyin@sdust.edu.cn

Abstract: This paper introduces a novel framework for zero-shot learning (ZSL), i.e., to recognize new categories that are unseen during training, by distilling knowledge from foundation models. Specifically, we first employ ChatGPT and DALL-E to synthesize reference images of unseen categories from text prompts. Then, the test image is aligned with text and reference images using CLIP and DINO. Finally, the predicted logits are aggregated according to their confidence to produce the final prediction. Experiments are conducted on multiple datasets, including CIFAR-10, CIFAR-100, and TinyImageNet. The results demonstrate that our model can significantly improve classification accuracy compared to previous approaches, achieving AUROC scores above 96% across all test datasets. Our code is available at https://github.com/1134112149/MICW-ZIC.

Keywords: zero-shot learning; multi-method integration; image classification

1. Introduction

In recent years, deep learning technologies have made significant progresses, particularly in the field of image classification, with numerous successful models being developed. For instance, ResNet [1] introduced a residual learning mechanism to address the difficulties in training deep networks, substantially increasing classification accuracy. Meanwhile, MobileNet [2] achieved efficient computation on mobile and embedded devices with its lightweight deep separable convolution design. Despite the significant success, these methods are built upon an assumption that the categories of the test images have been seen in the training set, which imposes restrictions in their practical applications. In real-world scenarios, we often encounter a large number of new categories that may not be included in the original training set.

To overcome the limitations of the aforementioned methods when dealing with unseen categories, Zero-Shot Learning (ZSL) has garnered increasing attention in recent years. ZSL aims to enable models to recognize categories that they have never seen during the training phase, thus being able to identifying new classes by leveraging the semantic relationships between categories. Several prominent ZSL models have been developed from various aspects. For example, an early ZSL model [3] improved the classification accuracy of unseen bird species by learning the mapping between visual features and textual descriptions. Due to the relatively small number of images for training models, these methods still face challenges in achieving satisfactory zero-shot generalization capabilities.

Recently, CLIP (Contrastive Language-Image Pre-Training) is developed by leveraging the synergy between vision and language. Through extensive contrastive learning with large-scale image and text data, CLIP learns a universal visual-linguistic model, enabling direct zero-shot classification for new categories by aligning image features with textual descriptions. Specifically, CLIP evaluates the similarity between an image and a series of candidate text descriptions, allowing it to classify unseen categories without the need for additional training data. However, using text-image alignment for Zero-Shot Classification still poses challenges, such as semantic gap issues and contextual ambiguity problems, which limit the model's performance on complex or fine-grained categories.

In this paper, we propose a novel framework to leverage the alignment of images with text as well as the alignment among images themselves to further improve the ZSL accuracy by leveraging multiple foundation models like ChatGPT, CLIP and DINO. Specifically, we first employ ChatGPT and DALL-E to generate images of unseen categories from text prompts. Then, Then, the test image is aligned with

 doi:10.20944/preprints202406.1153.v1

text and reference images using CLIP and DINO. Finally, the predicted logits are aggregated according to their confidence to produce the final prediction. To evaluate the effectiveness of our method, we conduct experiments on multiple datasets including CIFAR-10, CIFAR-100, and TinyImageNet. In comparison to ZSL methods, our model has realized significant gains in classification accuracy, with improvements of 11.97%, 26.06%, and 12.6% on CIFAR-10; 24.48%, 42.48%, and 5.42% on CIFAR-100; and 32.32%, 41.96%, and 1.2% on TinyImageNet, respectively. Furthermore, our model score above 96% on the AUROC indicator across all test datasets, particularly exceeding 99% on the CIFAR-10 dataset. These results clearly demonstrate the superiority of our method in ZSL.

2. Related Work

2.1. Zero-Shot Classification

Zero-shot classification (ZSC) is a machine learning paradigm aimed at enabling models to recognize categories they have not seen during the training phase. Unlike traditional supervised learning methods, ZSC does not rely on direct experience with every class but rather achieves classification through understanding shared knowledge or attributes among categories. This capability is crucial for dealing with data scarcity or category diversity issues, especially in fields such as Natural Language Processing (NLP) and Computer Vision (CV). In recent years, significant progress has been made in the field of ZSC. In particular, pre-trained language models such as GPT-3 [4] and BERT [5] show great potential in ZSC tasks. These models, having been pre-trained on massive text datasets, learn rich linguistic features and world knowledge, enabling them to infer unknown categories without explicit examples. Additionally, by operating on graph-structured data, GNNs is able to capture complex relationships and attributes between categories, offering new avenues for accurate classification of unseen classes [6,7]. To further improve ZSL performance, Deep Calibration Networks (DCN) [8] aim to reduce the uncertainty of models when dealing with unseen categories. In addition, meta-learning has been studied to facilitate models to achieve higher accuracy on unseen images [9]. A framework based on Generative Adversarial Networks (GANs) is proposed in [10], which can generate features representative of unseen categories to achieve further gains. To address the data imbalance issue in ZSL, Balanced Meta-Softmax [11] is developed to achieve a better balance between different categories.

With the rapid advancements of recent foundation models, numerous works have been made to study these models from the perspective of ZSL. CLIP [12] has its superior performance across multiple computer vision benchmarks, particularly showing its powerful capability in zero-shot classification tasks. It highlights the potential of learning visual models with natural language supervision. Several attempts [13–15] have been made to adopt CLIP to classify images of unseen categories. DINO [16] has demonstrated the superior generalization capability of its features in downstream tasks like image classification. A k-NN classifier is employed in [16] to directly leverage self-supervised ViT features for subsequent semantic segmentation, which show great potential for ZSC. DALL-E is a powerful image generator that can synthesize images from a given text prompt. Zhang *et al.* [17] employ generative models to produce additional images to augment the training dataset to achieve gains in ZSL.

However, the mechanism of our method is significantly different. We utilize the extensive knowledge of ChatGPT and the powerful image generation capabilities of DALL-E to create reference images that precisely describe unseen categories and classification boundaries, obtaining reference samples for new categories and those at the classification boundaries. This differs from the approach by [17], which aimed to augment training samples using DALL-E. Unlike the preprocessing done during training in previous methods, our approach supplements high-quality boundary images during prediction as reference images. Additionally, we propose leveraging CLIP's image feature extraction capability, integrating not only its text-image alignment but also image-image alignment techniques during classification. Compared to models without this technique, ours achieved a respective increase of 0.45%, 0.24%, and 0.23% in classification accuracy on the CIFAR10, CIFAR100, and TinyImageNet datasets. Furthermore, we introduce a novel confidence-based adaptive weighting mechanism to

 doi:10.20944/preprints202406.1153.v1

aggregate results from different prediction models that is not utilized in [17]. This mechanism allows us to weigh the results of each prediction model based on its confidence level, enabling more effective integration of multiple model predictions. Such adaptive weighting enhances the flexibility of our approach to accommodate varying performances of different models in diverse scenarios, further improving overall prediction accuracy and robustness.

3. Method

The overview of our method is summarized in Figure 1. Assume we have a dataset (D) containing m seen classes and n unseen classes. When implementing the classification task of $m + n$ (defined as N) categories within the dataset D, we first generate some reference images for each category. During the test phase, the test image x_t is first aligned with class-definition text using CLIP. Then, the test image is also aligned with reference images using CLIP and DINO. Finally, the resultant logits are aggregated according to their confidence to obtain the final prediction.

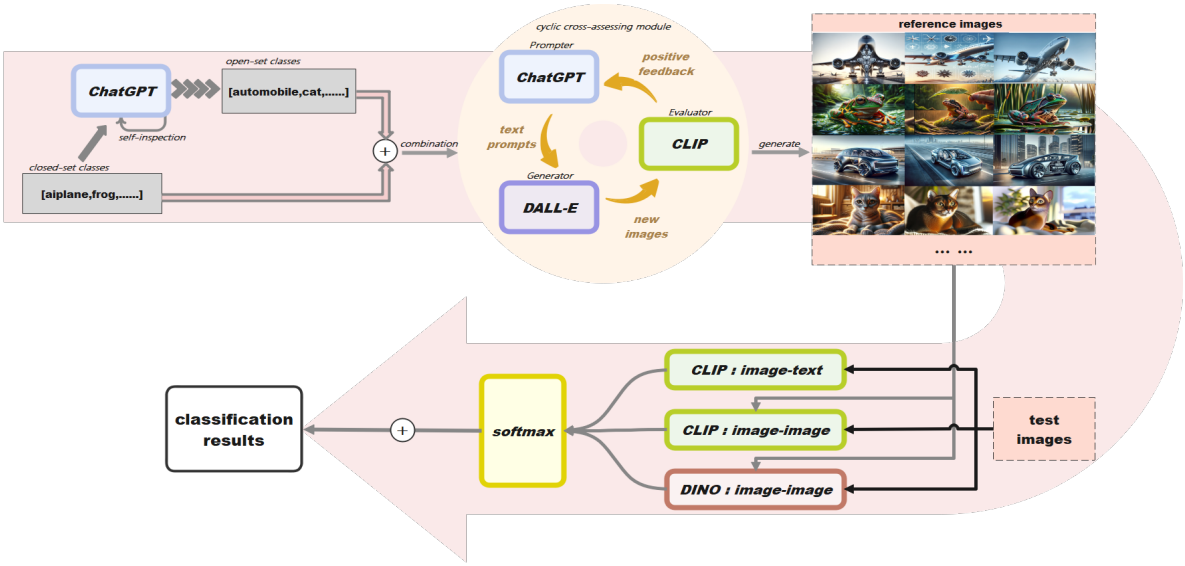

Figure 1. Overview of our method.

3.1. Reference Images Generation

To synthesize images of unseen categories, ChatGPT and DALL-E are employed to leverage their rich knowledge. Note that, several categories are easily confused in appearance since they may be similar in shape or features. Thus, we propose to generate reference images that are located near the boundaries of ambiguous categories. Specifically, we first conduct feature analysis to obtain common features of similar categories. Then, we synthesize images based on the resultant common features.

(1) Feature Analysis

First, ChatGPT is adopted to analyze the appearance of N categories in the dataset, as shown in the left column of Figure 2. Then, ChatGPT is used to group categories that are similar in appearance. Specifically, we ask ChatGPT to identify those similar categories that are easily confused. Next, for each pair of resultant groups, ChatGPT is further adopted to analyze their common appearance features. Finally, the common features are obtained and denoted as cc.

 doi:10.20944/preprints202406.1153.v1

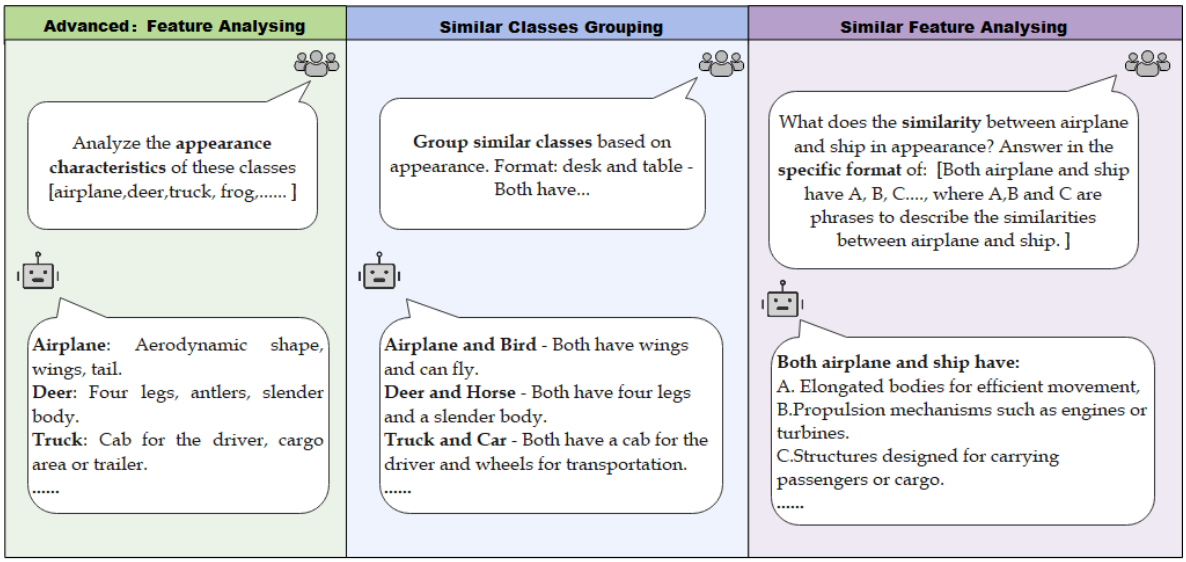

Figure 2. Examples of utilizing ChatGPT to generate common features of similar categories.

(2) Image Synthesis

After feature analysis, DALL-E is employed to generate images of each class that possess the common appearance features cc. For categories that do not resemble other categories, we directly utilize DALL-E to generate images belonging to it without using detailed appearance descriptions. To acquire more comprehensive reference information, we generate multiple images for each category. All these images will be employed in the subsequent steps for image-image alignment.

3.2. Alignment-Driven Image Classification

To achieve accurate classification of an unseen test image, we align the image with text prompt and reference images using foundation models to distill their knowledge.

3.2.1. Text-Image Alignment

CLIP is capable of encoding image and text into a shared representational space. With this property, the logits of the test image x_t and the unseen category c can be obtained by encoding x_t and c using CLIP and calculating the cosine similarity S of the resultant representations:

$$S^n_{i-t,clip} = \mathrm{Cos}(F_i(x_t), F_t(T_n)). \tag{1}$$

3.2.2. Image-Image Alignment

In addition to text-image alignment, we also conduct image-image alignment to leverage powerfull vision foundation models for ZSC. Specifically, the test image and the reference images are first encoded using CLIP and DINO to obtain their representations and then calculate their cosine similarities.

$$S^n_{i-i,clip} = \frac{1}{M}\sum_{i=1}^{M} \mathrm{Cos}\Big(F_i(x_t), F_i(r^i_n)\Big). \tag{2}$$

$$S^n_{i-i,dino} = \frac{1}{M}\sum_{i=1}^{M} \mathrm{Cos}\Big(D(x_t), D(r^i_n)\Big). \tag{3}$$

where D refers to the image encoder of DINO.

 doi:10.20944/preprints202406.1153.v1

3.3. Logits Aggregation

After text-image and image-image alignment, a confidence-based strategy is employed to aggregate the resultant logits. That is, the logits with higher confidence are assigned with larger weights to produce the final results. Next, the final output is obtained as:

$$S_{fuse} = W_{i-t,clip} \cdot S_{i-t,clip} + W_{i-i,clip} \cdot S_{i-i,clip} + W_{i-i,dino} \cdot S_{i-i,dino}. \quad (4)$$

4. Experiments

4.1. Datasets and Evaluation Metrics

To evaluate the effectiveness of our proposed method, we conduct experiments on three widely-used datasets, including CIFAR10, CIFAR100, and TinyImageNet. For CIFAR10, we select 6 categories as the closed set and the remaining 4 categories as the open set. For CIFAR10, 10 categories are chosen as the closed set, with the remaining 90 categories serving as the open set. For TinyImageNet, we select 20 categories as the closed set, with the other 180 categories as the open set. We employ Accuracy (AC) and Area Under the Receiver Operating Characteristic curve (AUROC). The AUROC metric is used to evaluate the model's ability to differentiate between closed-set and open-set samples, that is, the accuracy of the model in identifying novel category samples.

4.2. Main Results

We compare our method with nine state-of-the-art methods, including RPL [18], OpenHybrid [19], PMAL [20], ZOC [21], MLS [22], DIAS [23], Class-iinclusion [24], ODL [25], and CSSR [26]. Quantitative results are presented in Table 1. Compared with previous open-set classification methods, our method produces significant accuracy gains, especially on CIFAR-10 and TinyImageNet. For example, our method outperforms Class-inclusion by over 5% and 3% on CIFAR-10 and TinyImageNet, respectively. The higher accuracy of our method clearly demonstrate its effectiveness and superiority.

Table 1. AUROC results achieved by different methods.

Method	Venue	CIFAR10	CIFAR+10	TinyImageNet
Methods that involve a training process				
RPL [18]	ECCV 2020	90.1	97.6	80.9
OpenHybrid [19]	ECCV 2020	95.0	96.2	79.3
PMAL [20]	AAAI 2022	95.1	97.8	83.1
ZOC [21]	AAAI 2022	93.0	97.8	84.6
MLS [22]	ICLR 2022	93.6	97.9	83.0
DIAS [23]	ECCV 2022	85.0	92.0	73.1
Class-inclusion [24]	ECCV 2022	94.8	96.1	78.5
ODL [25]	TPAMI 2022	85.7	89.1	76.4
ODL+ [25]	TPAMI 2022	88.5	91.1	74.6
CSSR [26]	TPAMI 2022	91.3	96.3	82.3
RCSSR [26]	TPAMI 2022	91.5	96.0	81.9
Methods that involve no extra training process				
Ours		**99.78**	96.03	**96.48**

4.3. Ablation Study

4.3.1. Effectiveness of Different Components

We first conduct experiments to study the effectiveness for different components of our method, including image-text alignment using CLIP (M1), image-image alignment using CLIP (M2)m and image-image alignment using DINO (M3). Specifically, we compare the performance of our method with different combinations of components. As shown in Table 2 and Table 3, all the components in

 doi:10.20944/preprints202406.1153.v1

our methods contribute to higher accuracy. Without M1/M2/M3, the network variant suffers notable accuracy drop on most metrics. Moreover, M1 and M3 contribute more significantly than M2. These results clearly demonstrate the effectiveness of each component in our method.

Table 2. Top1, Top3, Top5 results achieved by our method with different settings.

Method	CIFAR10			CIFAR100			TinyImageNet		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
M1+M2	83.83%	94.85%	98.22%	48.00%	66.41%	73.27%	43.52%	63.97%	71.79%
M1+M3	92.51%	98.16%	**99.35%**	71.93%	**86.59%**	**90.50%**	73.29%	84.73%	88.41%
M2+M3	79.99%	93.22%	97.41%	64.19%	79.29%	84.63%	72.74%	84.33%	87.30%
Ours (M1+M2+M3)	**92.96%**	**98.32%**	99.33%	**72.17%**	86.55%	90.36%	**73.52%**	**84.95%**	**88.55%**

Table 3. AUROC results achieved by our method with different settings.

Method	CIFAR10	CIFAR100	TinyImageNet
M1+M2	97.94%	92.99%	87.03%
M1+M3	99.75%	95.87%	**96.54%**
M2+M3	97.26%	94.70%	96.09%
Ours (M1+M2+M3)	**99.78%**	**96.03%**	96.48%

4.3.2. Effects of Different Aggregation Strategies

We further conduct experiments to study different aggregation strategies. Specifically, we first use predefined weights to aggregate the logits produced by different methods. Then, we study three different weighting strategies, which employ max similarity, inverse entropy, and negative exponential of entropy to fuse logits. As shown in Table 4 and 5, the method with inverse entropy produces the best performance. Consequently, this strategy is used as the default setting in our experiments.

Table 4. Top1, Top3, Top5 accuracy achieved by our method with different weight settings on CIFAR10.

Weighting Method	Top1	Top3	Top5
1:1:1	92.36%	98.22%	99.31%
3:3:4	92.29%	98.21%	99.26%
Max Similarity	92.75%	98.26%	99.34%
Inverse Entropy	**92.96%**	**98.32%**	99.33%
Negative Exponential of Entropy	92.90%	98.31%	**99.36%**

Table 5. AUROC results achieved by our method with different weight setting methods on CIFAR10.

Weighting Method	AUROC
1:1:1	99.55%
3:3:4	99.60%
Max Similarity	99.68%
Inverse Entropy	**99.78%**
Negative Exponential of Entropy	99.73%

4.3.3. Number of Reference Images

In our method, the number of generated reference images has a huge impact on the ability to distinguish between different categories. Therefore, we conduct experiments to study the effects of the number of reference images. Specifically, we construct a network variant with only one reference image for a category and compare its accuracy with our baseline. As shown in Table 6, the model with

 doi:10.20944/preprints202406.1153.v1

multiple images outperforms the counterpart with one reference image significantly. This is because, more reference images can provide more information to better describe the properties of a category, which ultimately contributes to higher accuracy.

Table 6. Comparison of Top1 and AUROC results when using one and multiple reference images. The left side of each '-' represents the Top1 result, and the right side of each '-' represents the AUROC result.

Number of Reference Images	CIFAR10	CIFAR100	TinyImageNet
One	88.71%-99.05%	67.78%-96.025%	67.77%-96.00%
Multiple	**92.96%-99.78%**	**72.17%-96.026%**	**73.52%-96.48%**

4.3.4. Visualization of Reference Images

We also visualize the synthetic images in Figure 3. It can be observed that our method is able to synthesize images for similar categories with similar appearance, which facilitate our method to produce higher accuracy.

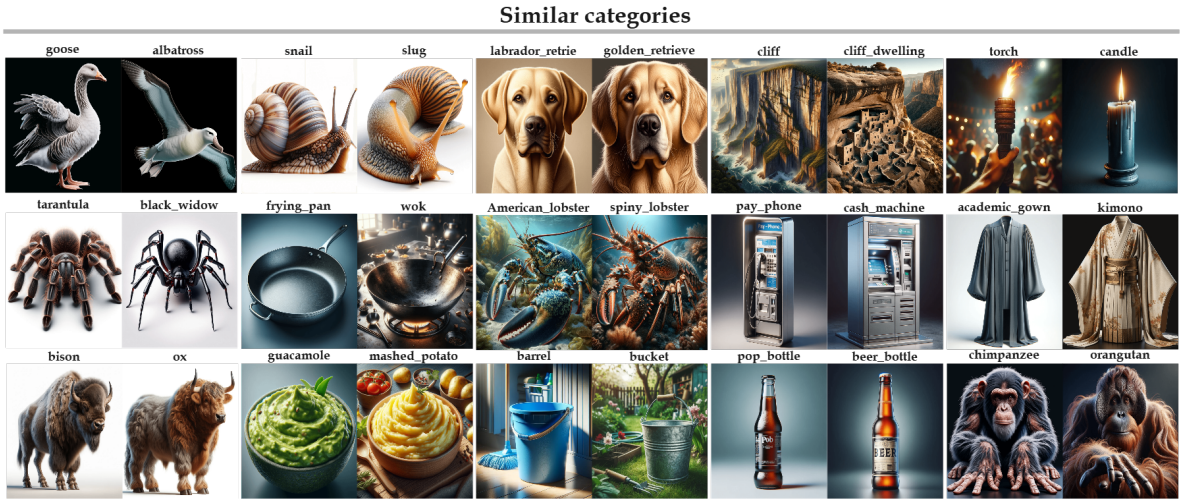

Figure 3. Examples of synthesized reference images.

5. Conclusion

In this paper, we introduce a framework for zero-shot classification by exploiting knowledge from multiple foundation models. We first employ ChatGPT and DALL-E to synthesize reference image of unseen categories. Then, text-image alignment and image-image alignment are conducted to produce the logits of the test image. Finally, the resultant logits are aggregated to generate the overall prediction. Experimental results on CIFAR-10, CIFAR-100, and TinyImageNet validate the effectiveness of our approach, showcasing its exceptional performance in both open and closed set classification tasks.

References

1. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
2. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861* **2017**.
3. Paz-Argaman, T.; Atzmon, Y.; Chechik, G.; Tsarfaty, R. Zest: Zero-Shot Learning from Text Descriptions Using Textual Similarity and Visual Summarization. *arXiv preprint arXiv:2010.03276* **2020**.
4. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; others. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.

 doi:10.20944/preprints202406.1153.v1

5. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.
6. Ou, G.; Yu, G.; Domeniconi, C.; Lu, X.; Zhang, X. Multi-label zero-shot learning with graph convolutional networks. *Neural Networks* **2020**, *132*, 333–341.
7. Gao, J.; Xu, C.S. CI-GNN: Building a category-instance graph for zero-shot video classification. *IEEE Transactions on Multimedia* **2020**, *22*, 3088–3100.
8. Liu, S.C.; Long, M.S.; Wang, J.M.; Jordan, M.I. Generalized zero-shot learning with deep calibration network. *Advances in neural information processing systems* **2018**, *31*.
9. Sankaranarayanan, S.; Balaji, Y. Meta learning for domain generalization. In *Meta Learning With Medical Imaging and Health Informatics Applications*; Elsevier, 2023; pp. 75–86.
10. Xian, Y.Q.; Lorenz, T.; Schiele, B.; Akata, Z. Feature generating networks for zero-shot learning. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5542–5551.
11. Ren, J.W.; Yu, C.J.; Ma, X.; Zhao, H.Y.; Yi, S.; others. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems* **2020**, *33*, 4175–4186.
12. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; others. Learning transferable visual models from natural language supervision. International conference on machine learning. PMLR, 2021, pp. 8748–8763.
13. Shipard, J.; Wiliem, A.; Thanh, K.N.; Xiang, W.; Fookes, C. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 769–778.
14. Christensen, A.; Mancini, M.; Koepke, A.; Winther, O.; Akata, Z. Image-free Classifier Injection for Zero-Shot Classification. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 19072–19081.
15. Novack, Z.; McAuley, J.; Lipton, Z.C.; Garg, S. Chils: Zero-shot image classification with hierarchical label sets. International Conference on Machine Learning. PMLR, 2023, pp. 26342–26362.
16. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 9650–9660.
17. Zhang, R.; Hu, X.; Li, B.; Huang, S.; Deng, H.; Qiao, Y.; Gao, P.; Li, H. Prompt, Generate, then Cache: Cascade of Foundation Models Makes Strong Few-Shot Learners. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15211–15222.
18. Chen, G.; Qiao, L.; Shi, Y.; Peng, P.; Li, J.; Huang, T.; Pu, S.; Tian, Y. Learning Open Set Network with Discriminative Reciprocal Points. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III. Springer, 2020, pp. 507–522.
19. Zhang, H.; Li, A.; Guo, J.; Guo, Y. Hybrid Models for Open Set Recognition. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III. Springer, 2020, pp. 102–117.
20. Lu, J.; Xu, Y.; Li, H.; Cheng, Z.; Niu, Y. PMAL: Open Set Recognition via Robust Prototype Mining. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 1872–1880.
21. Esmaeilpour, S.; Liu, B.; Robertson, E.; Shu, L. Zero-shot Out-of-Distribution Detection Based on the Pre-trained Model CLIP. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 6568–6576.
22. Vaze, S.; Han, K.; Vedaldi, A.; Zisserman, A. Open-Set Recognition: A Good Closed-Set Classifier Is All You Need? **2021**.
23. Moon, W.; Park, J.; Seong, H.S.; Cho, C.H.; Heo, J.P. Difficulty-Aware Simulator for Open Set Recognition. European Conference on Computer Vision. Springer, 2022, pp. 365–381.
24. Cho, W.; Choo, J. Towards Accurate Open-Set Recognition via Background-Class Regularization. European Conference on Computer Vision. Springer, 2022, pp. 658–674.

 doi:10.20944/preprints202406.1153.v1

25. Liu, Z.g.; Fu, Y.m.; Pan, Q.; Zhang, Z.w. Orientational Distribution Learning with Hierarchical Spatial Attention for Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**.
26. Huang, H.; Wang, Y.; Hu, Q.; Cheng, M.M. Class-Specific Semantic Reconstruction for Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, *45*, 4214–4228.