

MOS: Modeling Object-Scene Associations in Generalized Category Discovery

Zhengyuan Peng¹, Jinpeng Ma³, Zhimin Sun¹, Ran Yi¹, Haichuan Song², Xin Tan^{2,*†}, Lizhuang Ma^{1,2,*}
¹Shanghai Jiao Tong University, ²East China Normal University, ³Chongqing University

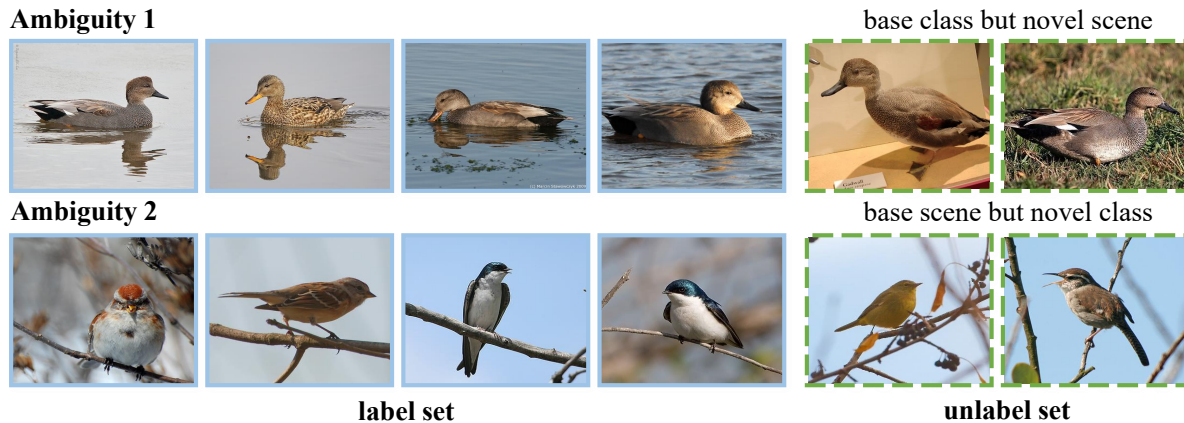


Figure 1. **Illustration of the ambiguity challenge.** The top row illustrates the potential for base objects in novel scenes to be perceived as novel categories. The bottom row shows misclassification risks when novel objects are placed in base scenes. Ambiguity challenge is the primary factor that leads to the misinterpretation of scene information.

Abstract

Generalized Category Discovery (GCD) is a classification task that aims to classify both base and novel classes in unlabeled images, using knowledge from a labeled dataset. In GCD, previous research overlooks scene information or treats it as noise, reducing its impact during model training. However, in this paper, we argue that scene information should be viewed as a strong prior for inferring novel classes. We attribute the misinterpretation of scene information to a key factor: the **Ambiguity Challenge** inherent in GCD. Specifically, novel objects in base scenes might be wrongly classified into base categories, while base objects in novel scenes might be mistakenly recognized as novel categories. Once the ambiguity challenge is addressed, scene information can reach its full potential, significantly enhancing the performance of GCD models. To more effectively leverage scene information, we propose the **Modeling Object-Scene Associations (MOS)** framework, which utilizes a simple MLP-based scene-awareness module to enhance GCD perfor-

mance. It achieves an exceptional average accuracy improvement of **4%** on the challenging fine-grained datasets compared to state-of-the-art methods, emphasizing its superior performance in fine-grained GCD. The code is publicly available at <https://github.com/JethroPeng/MOS>.

1. Introduction

Deep learning models in image recognition often face challenges when novel classes are introduced, due to their reliance on large annotated datasets and closed-world assumption. This limitation sparks interest in Generalized Category Discovery (GCD) [42], which aims to classify novel categories from unlabeled data using knowledge from labeled datasets. GCD enables models to generalize to novel categories and adapt to diverse environments, overcoming the constraints of the closed-world assumption.

In GCD, existing approaches [3, 10, 12, 14, 16, 28, 30, 33, 42, 43, 45, 48, 52, 53, 55–58] often overlook scene information (e.g., forest, ocean, sky) or treat it as noise that interferes with the generalization of classification models, seeking to minimize its influence through various techniques. This viewpoint has long been widely accepted, sup-

*Corresponding author.

†Project Leader.

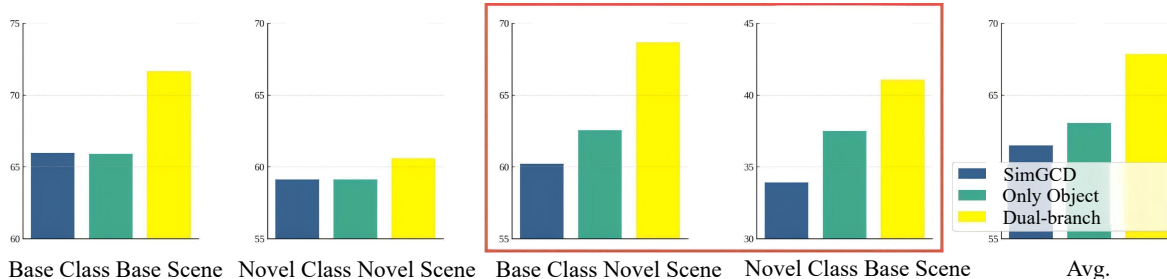


Figure 2. **Comparison of the performance on the 4 subsets of the CUB dataset.** The red box highlights scenarios involving two types of ambiguities. We compare the performance variations between the scene-removed image (green) and the original image (dark blue) as inputs. Our analysis reveals that the most significant performance improvement occurs in situations involving a novel-base relationship conflict between object and scene. In other situations, there is a slight decline in performance. Furthermore, employing dual-branch network for training (yellow) leads to a notable performance enhancement across four subsets.

ported by a key experiment: performance notably improves in GCD when the scene is removed. However, we argue that the scene is not noise. Instead, in environments with limited information about novel classes, scene information can actually serve as a useful signal for inferring the class, providing strong prior knowledge that supports more accurate predictions.

This paper finds that scene information is often misinterpreted as noise due to **Ambiguity Challenge** inherent in GCD. Specifically, when a novel scene contains base objects, these objects are more likely to be misclassified as novel due to the difficulty in distinguishing between scene and object. Conversely, novel objects within a familiar scene may be erroneously classified into base categories. We conduct an observation experiment on the CUB dataset. In our experiments, we categorize scenes as novel or base class based on their presence in the labeled set. Based on the base and novel relationships between objects and scenes, we divide the dataset into four subsets. We find that the improvement from scene removal primarily occurs in situations where the base-novel relationship between the object and the scene conflict, as is shown in Fig. 2. We illustrate ambiguity challenge in Fig. 1, which confirms our views.

Further experiments demonstrate that once the ambiguity challenge is addressed, the scene information can enhance object classification performance. Based on the scene-removed image and original image, we further employ a dual-branch network for joint learning, where the model can learn scene information through the contrastive differences. Compared to the performance of object image, results suggest that enabling the network to learn scene information achieves significant improvements in accuracy, outperforming the single-branch model across all four subsets. Results are shown in Fig. 2. In Sec. 4.6, we investigate whether the network effectively captures the scene information.

To more effectively leverage scene information, we propose the **Modeling Object-Scene Associations (MOS)**

framework, which utilizes a simple MLP-based scene-awareness module to enhance GCD performance. In our framework, we employ a universal saliency segmentation model to perform zero-shot segmentation of objects and scenes. Their features are then extracted in the same feature space using a shared backbone for consistent representation. Scene-awareness module processes these two features for perceptual differentiation, extracting more effective criteria for categorization. Following extensive experimentation, MOS surpasses state-of-the-art in fine-grained GCD tasks, with only a minimal increase in network. Notably, it achieves an impressive average accuracy of **65%** on three parts of the Semantic Shift Benchmark (SSB), underlining its superior performance. Compared to the baseline, we achieve a **9%** performance improvement.

Our research indicates that scene information provides more benefits than drawbacks in fine-grained GCD task. Reducing scene information introduces a performance bottleneck in scenarios with limited information. Effective mining of scene information can provide more comprehensive criteria for discovering different novel classes.

We summarize our contributions as follows:

1. We challenge the traditional view that scene information interferes with GCD. Our experiments show that, when the ambiguity challenge is addressed, scene information can serve as a valuable prior for inferring novel categories, improving fine-grained GCD performance.
2. We propose a novel framework, **Modeling Object-Scene Associations (MOS)**, which incorporates a simple MLP-based scene-awareness module to effectively leverage scene information in GCD. Our approach significantly outperforms existing state-of-the-art methods on fine-grained datasets.
3. We annotate the scene information in the CUB dataset to enable further exploration of scene impacts and support related research.

2. Related work

Semi-Supervised Learning leverages both labeled and unlabeled data to enhance model training, proving particularly valuable in situations where only a small proportion of the dataset is labeled. Basic methods in semi-supervised learning include self-training [19, 50], consistency regularization [18, 41], and other approaches [2, 6, 8, 9, 13, 35]. Self-training consists of two stages: initial training on labeled data, followed by further training with pseudo-labels generated from unlabeled data. Consistency regularization ensures that predictions for augmented versions of the same data remain consistent. Another important variant of the semi-supervised learning is open-set semi-supervised learning [7, 22, 47], which focuses on discovering outliers in unlabeled datasets that do not correspond to any category in the labeled set. However, it generally does not distinguish between different types of outliers.

Generalized Category Discovery aims to classify both base and novel classes in unlabeled images, using knowledge from a labeled dataset. A seminal study [42] fine-tunes DINO features and categorizes different classes using semi-supervised k-means. SimGCD [52] improves upon this by employing a parametric model instead of clustering algorithms, enhancing performance and robustness. DCCL [30] introduces an innovative iterative framework that simultaneously estimates underlying visual concepts and learns their representations. PromptCAL [55] proposes a two-stage Contrastive Affinity Learning method using visual prompts. SPTNet [45] employs a two-stage approach that iteratively fine-tunes both models and input data. In addition to these approaches, other methods [3, 10–12, 14, 16, 24, 28, 29, 31, 33, 34, 43, 48, 53, 56–59] also improve the performance of GCD from different perspectives. Most of them overlook scene information or treat it as noise. However, the underlying mechanisms and role of scene information in GCD remain underexplored.

Scene Information in classification. Scene is an important topic in computer vision [15, 20, 21, 23, 36–40, 46, 51, 54]. The influence of scene information on classification performance remains controversial. Some studies [54] suggest that scene information primarily introduces spurious correlations, which can lead to noise and hinder the generalization of classification models, rather than contributing to meaningful classification signals. To this end, some models [5, 26] are designed to prioritize object-centric information, utilizing attention modules that focus predominantly on objects. Concurrently, there is an emerging recognition of the significance of contextual scene information in enhancing classification performance, supported by recent research [46]. It shows that scene elements in an image provide crucial prior knowledge that aids in classification. The above studies are typically based on traditional classification tasks. In GCD task, the relationships become more

complex with novel class. This paper attempts to delve into the impact of scene information in GCD.

3. Methodology

3.1. Preliminary and Overview

In Generalized Category Discovery (GCD), the dataset D comprises both a labeled subset D_l and an unlabeled subset D_u , defined as $D = D_l \cup D_u = \{(X, Y)\}$. The primary goal is to use the labeled dataset $D_l = \{(X_l, Y_l)\}$ within the known label space C_l , to effectively categorize samples in the unlabeled dataset $D_u = \{(X_u)\}$. The label space of these unlabeled samples, C_u , often includes C_l as a subset, $C_l \subset C_u$. This setup presents a complex classification challenge as C_u includes additional categories unseen in C_l .

In this section, we provide a detailed illustration of our method, **Modeling Object-Scene Associations (MOS)**. MOS takes two inputs: the original image X and the segmented object O . In Sec. 3.2, we first define the method for object extraction. Sec. 3.3 offers an overview of the MOS framework and describes its algorithmic flow. Sec. 3.4 delves into the scene-awareness module, a key component of the **MOS** framework, which enhances performance.

3.2. Object-Scene Decoupling

We utilize the universal zero-shot saliency segmentation model IS-Net [32] to separate objects from their scene. The segmentation process generates a saliency mask $M \in \{0, 1\}^{m \times n}$, where foreground objects are represented by 1 and the background by 0. We extract object image O from the original image X using the formula:

$$O = X \cdot M + \mu \cdot (1 - M), \quad (1)$$

where μ represents a filling value used to replace the scene pixels, m and n denote the width and height of the original image X , respectively. In practice, we set μ to the mean pixel value of the image. This method is essential for minimizing the domain gap caused by missing scene parts.

3.3. Modeling Object-Scene Associations

Scene information provides essential contextual knowledge for classification. Thus, we introduce the novel **MOS** framework, which effectively models object-scene associations with only a minimal increase in network. **MOS** framework includes two primary components: the backbone network f and the scene-awareness module θ . Compared to SimGCD [52], **MOS** introduces only minimal additional training parameters, specifically by adding a single training MLP in scene-awareness module.

The input to MOS consists of the original image and the image after object extraction. The method for object extraction is described in Sec. 3.2. The shared backbone network extracts both object and scene features simultaneously

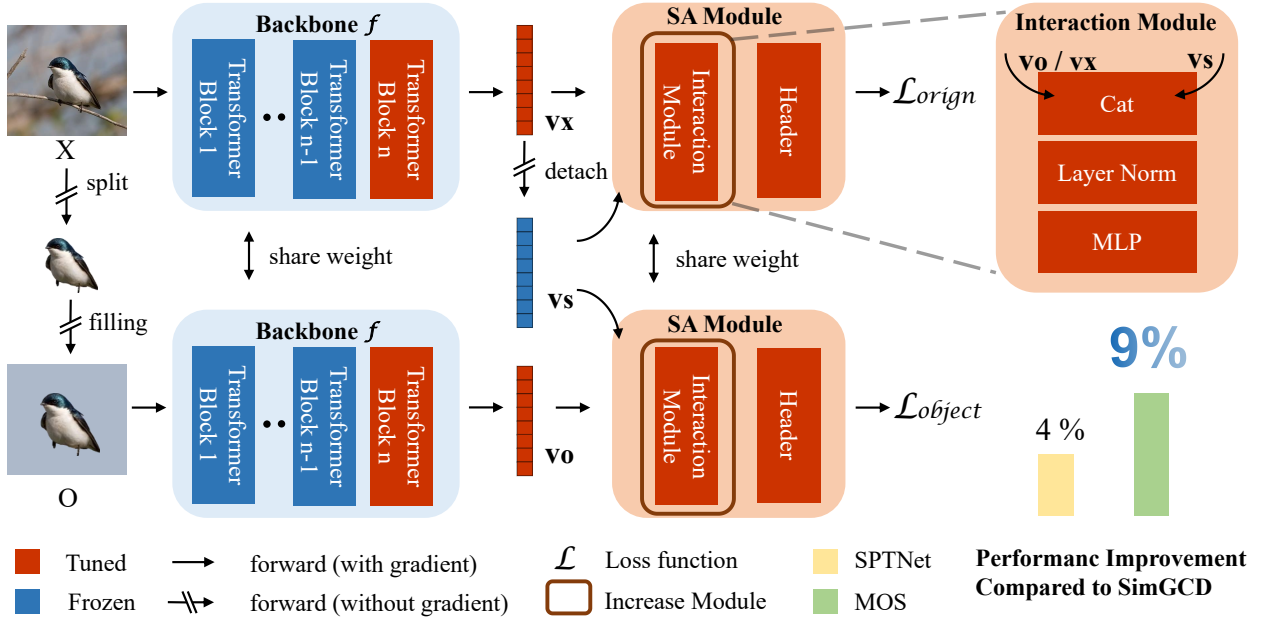


Figure 3. **Modeling Object-Scene Associations (MOS) Framework.** The framework adopts a dual-branch design: one branch processes the original image, and the other handles the segmented object image. Both branches share two core components: the **Backbone** f and the **Scene-awareness Module** θ . During training, the original image is segmented using a universal saliency segmentation model to extract the object. The scene regions of the object image are then filled with the mean pixel value. Both the original and object images are fed into the Backbone. After extracting features v_o from the object image and v_x from the original image, the scene features v_s are obtained from v_x . In the scene-awareness module, v_x and v_s , as well as v_o and v_s , interact to produce the classification output and compute the losses. During evaluation, only the output from the object branch is used.

within the same feature space. We use the information obtained from the object image as object features. As for the scene features, scene information S , typically located at the image edges, is difficult for the DINO network to extract as standalone features. Given that the conditional entropy of the scene features is consistent with the conditional entropy of the original image feature under the same condition of object features (The original image only consist of both scene and object components), we therefore replace the scene features with the original image features.

Our framework is divided into two branches: one processes the object parts O , and the other focuses on the original images X . The overall pipeline is listed as follows:

$$\left. \begin{array}{l} v_o = f(O) \\ v_x = f(X) \\ v_s := v_x \end{array} \right\} \text{Feature} \quad \left. \begin{array}{l} \hat{y}_{\text{origin}} = \theta(v_x, v_s) \\ \hat{y}_{\text{object}} = \theta(v_o, v_s) \end{array} \right\} \text{Output}, \quad (2)$$

where \hat{y}_{origin} and \hat{y}_{object} dictate the output of two branch. In the original image branch, the primary distinction from the object image branch includes using the image features v_x , instead of v_o . v_x and v_s are interacted within the scene-awareness module to produce the classification output.

Two branches share the same backbone f and the SA

module θ , but losses are calculated separately. For each branch, we compute losses for both labeled and unlabeled images. For labeled images, we calculate the supervised contrastive learning loss $L_{\text{sup}}^{\text{nce}}$ and the classification loss $L_{\text{sup}}^{\text{cls}}$. For unlabeled images, we assess the unsupervised contrastive learning loss $L_{\text{un}}^{\text{nce}}$ and teacher-student cross-entropy loss $L_{\text{un}}^{\text{cls}}$. $L_{\text{sup}}^{\text{nce}}$ and $L_{\text{un}}^{\text{nce}}$ pull same-class features closer and push different-class features in the feature space. $L_{\text{sup}}^{\text{cls}}$ computes the cross-entropy between the output and labels, while $L_{\text{un}}^{\text{cls}}$ enforces consistency regularization for unlabeled set. The combined loss [52] is given as follows:

$$L_{\text{origin}}/L_{\text{object}} = (1 - \lambda)(L_{\text{un}}^{\text{nce}} + L_{\text{un}}^{\text{cls}}) + \lambda(L_{\text{sup}}^{\text{nce}} + L_{\text{sup}}^{\text{cls}}), \quad (3)$$

where λ is a weighting factor that balances supervised and unsupervised losses. The overall loss L is given as follows:

$$L = \lambda_1 L_{\text{origin}} + \lambda_2 L_{\text{object}}, \quad (4)$$

where λ_1 and λ_2 are weighting factors that balance the influence of each branch. In practical applications, setting both λ_1 and λ_2 to 1 is a common strategy to ensure balanced training and help improve performance. During evaluation, both object and original images are input and processed by the backbone f to extract v_o , v_x , and v_s . v_o and v_s pass

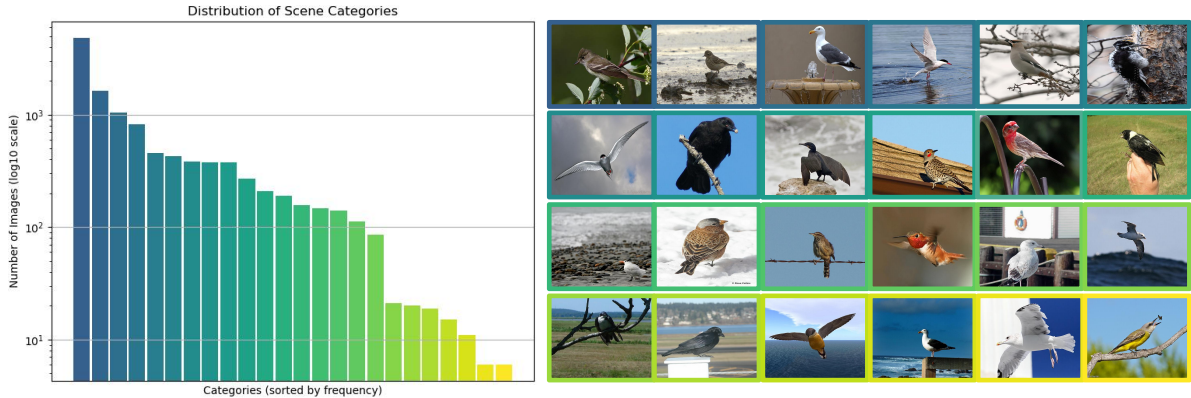


Figure 4. **Display of CUB Scene Information.** The left side shows category counts with a log-transformed y-axis to better illustrate the distribution. The right side shows category examples, illustrating the dataset’s diversity.

through object branch’s SA Module for output, while original image branch’s SA Module is discarded.

3.4. Scene-awareness Module

In **MOS**, the features from the original image X are denoted as v_x , and the features corresponding to the object O are represented as v_o . The scene features v_s are also derived from the original image features v_x . However, it introduces two challenges during training: 1) Replacing v_s with v_x leads to rapid variations in the feature space, making it difficult to capture meaningful information differences; 2) Over-optimization of v_x can cause an imbalance in the learning process, potentially compromising performance. To address these issues, we introduce a teacher network that remains fixed during training (i.e., The teacher output v_s is typically treated as a student output v_x with no gradient propagation (detach)). The teacher network outputs the scene features v_s , thereby maintaining the stability of the features and alleviating the challenges associated with training.

We input these features into the scene-awareness module, a concise and efficient interaction design. It only includes an Interaction Module IM and Header. IM normalizes the merged features to reduce discrepancies between inputs from different branches. We only need to employ a Multilayer Perceptron (MLP) for interaction. The formula for the Interaction Module IM is as follows:

$$IM_i = \text{MLP}\left(\frac{\mathbf{v}_i \oplus \mathbf{v}_s}{\|\mathbf{v}_i \oplus \mathbf{v}_s\|}\right), \quad i \in \{o, x\}, \quad (5)$$

where the feature vectors for the original image, object, and scene are denoted as \mathbf{v}_x , \mathbf{v}_o and \mathbf{v}_s , respectively. These vectors are concatenated and subsequently normalized by their combined norm. At the final stage of the scene-awareness module, we handle output through a DINO header. To ensure fairness in comparisons, we remove one MLP layer from the DINO header to compensate for the modules we

added. The final classification is determined using the Hungarian matching algorithm.

Table 1. **Dataset Statistics.**

| | Label | | Unlabel | |
|----------------------|-------|-------|---------|-------|
| | Num | Class | Num | Class |
| CUB [49] | 1.5K | 100 | 4.5K | 200 |
| Stanford Cars [17] | 2.0K | 98 | 6.1K | 196 |
| FGVC-Aircraft [25] | 1.7K | 50 | 5.0K | 50 |
| Oxford-IIIT Pet [44] | 0.9K | 19 | 2.7K | 37 |

4. Experiments

4.1. Datasets

We assess our approach using various fine-grained benchmarks tailored for image recognition. We employ the CUB [49], Stanford Cars [17], and FGVC-Aircraft [25] datasets, all part of the Semantic Shift Benchmark (SSB). For a more comprehensive evaluation, we also use Oxford-IIIT Pet [44]. The dataset division adheres to established protocols. We subsample $|C_l|$ seen (labeled) classes and follow the previous work [33, 42], selecting 50% of samples the datasets for the labeled set D_l , with the remaining samples constituting the unlabeled set D_u .

For our observation experiment, we annotate the scene information of the CUB [49] dataset. There are 24 different categories, which exhibit a clear long-tail distribution, as shown in Fig. 4. We categorize unlabeled CUB data into four subsets: Base Class with Base Scene, Novel Class with Base Scene, Base Class with Novel Scene, and Novel Class with Novel Scene. The division between novel class and base classes follows the settings of the SSB benchmark. For the definition of scenes, categories that are absent or appear very infrequently in the label set are designated as

Table 2. **Evaluation on the fine-grained datasets.** Values in bold indicate the top results.

| Method | Venue/Year | CUB | | | Stanford Cars | | | FGVC-Aircraft | | | Avg. | | |
|-----------------|------------|-------------|-------------|-------------|---------------|-------------|-------------|---------------|-------------|-------------|-------------|-------------|-------------|
| | | All | Base | Novel | All | Base | Novel | All | Base | Novel | All | Base | Novel |
| k-means++ [1] | SODA/2007 | 34.3 | 38.9 | 32.1 | 12.8 | 10.6 | 13.8 | 12.9 | 12.9 | 12.8 | 20.0 | 20.8 | 19.6 |
| RankStats+ [16] | TPAMI/2021 | 33.3 | 51.6 | 24.2 | 28.3 | 61.8 | 12.1 | 27.9 | 55.8 | 12.8 | 29.8 | 56.4 | 16.4 |
| UNO+ [14] | ICCV/2021 | 35.1 | 49.0 | 28.1 | 35.5 | 70.5 | 18.6 | 28.3 | 53.7 | 14.7 | 33.0 | 57.7 | 20.5 |
| GCD [42] | CVPR/2022 | 51.3 | 56.6 | 48.7 | 39.0 | 57.6 | 29.9 | 45.0 | 41.1 | 46.9 | 45.1 | 51.8 | 41.8 |
| XCon [12] | BMVC/2022 | 52.1 | 54.3 | 51.0 | 40.5 | 58.8 | 31.7 | 47.7 | 44.4 | 49.4 | 46.8 | 52.5 | 44.0 |
| ORCA [3] | ICLR/2022 | 36.3 | 43.8 | 32.6 | 31.9 | 42.2 | 26.9 | 31.6 | 32.0 | 31.4 | 33.3 | 39.3 | 30.3 |
| DCCL [30] | CVPR/2022 | 63.5 | 60.8 | 64.9 | 43.1 | 55.7 | 36.2 | - | - | - | - | - | - |
| GPC [58] | ICCV/2023 | 52.0 | 55.5 | 47.5 | 38.2 | 58.9 | 27.4 | 43.3 | 40.7 | 44.8 | 44.5 | 51.7 | 39.9 |
| PIM [10] | ICCV/2023 | 62.7 | 75.7 | 56.2 | 43.1 | 66.9 | 31.6 | - | - | - | - | - | - |
| SimGCD [52] | ICCV/2023 | 60.3 | 65.6 | 57.7 | 53.8 | 71.9 | 45.0 | 54.2 | 59.1 | 51.8 | 56.1 | 65.5 | 51.5 |
| PromptCAL [55] | CVPR/2023 | 62.9 | 64.4 | 62.1 | 50.2 | 70.1 | 40.6 | 52.2 | 52.2 | 52.3 | 55.1 | 62.2 | 51.7 |
| InfoSieve [33] | NIPS/2023 | 69.4 | 77.9 | 65.2 | 55.7 | 74.8 | 46.4 | 56.3 | 63.7 | 52.5 | 60.5 | 72.1 | 54.7 |
| μ GCD [43] | NIPS/2023 | 65.7 | 68.0 | 64.6 | 56.5 | 68.1 | 50.9 | 53.8 | 55.4 | 53.0 | 58.7 | 63.8 | 56.2 |
| GCA [27] | WACV/2024 | 68.8 | 73.4 | 66.6 | 54.4 | 72.1 | 45.8 | 52.0 | 57.1 | 49.5 | 58.4 | 67.5 | 54.0 |
| SPTNet [45] | ICLR/2024 | 65.8 | 68.8 | 65.1 | 59.0 | 79.2 | 49.3 | 59.3 | 61.8 | 58.1 | 61.4 | 69.9 | 57.5 |
| LeGCD [4] | CVPR/2024 | 63.8 | 71.9 | 59.8 | 57.3 | 75.7 | 48.4 | 55.0 | 61.5 | 51.7 | 58.7 | 69.7 | 53.3 |
| MOS (our) | | 69.6 | 72.3 | 68.2 | 64.6 | 80.9 | 56.7 | 61.1 | 66.9 | 58.2 | 65.1 | 73.4 | 61.0 |

novel scenes, while all others are considered base scenes. This dataset can also be used for other tasks related to scene studies.

4.2. Evaluation Protocol

We assess model performance using clustering accuracy (ACC), aligning with standard benchmarks. This involves matching ground truth labels, y_i , with predicted labels, \hat{y}_i , and calculating the ACC as follows:

$$\text{ACC} = \frac{1}{|D_u|} \sum_{i=1}^{|D_u|} \mathbf{1}(y_i = G(\hat{y}_i)), \quad (6)$$

where G represents the optimal permutation that aligns the predicted labels with the ground truth most accurately.

4.3. Implementation Details

We propose the **MOS** framework, which builds upon the SimGCD baseline and incorporates a pre-trained ViT-B/16 backbone for robust feature extraction. Our data augmentation and parameter learning strategies follow those of prior studies [45, 52], ensuring consistency and comparability in performance evaluations. Specifically, our data augmentation pipeline includes scaling, horizontal flipping, cropping, and color jittering, which introduces a variety of training samples. For saliency segmentation, we use the IS-Net [32] model, a universal zero-shot saliency segmentation framework. We train with a batch size of 128 for 200 epochs, starting with an initial learning rate of 0.1, which decays according to a cosine schedule for each dataset. In addition,

Table 3. **Evaluation on the Oxford-IIIT Pet datasets.** Values in bold indicate the top results.

| Method | Venue/Year | All | Base | Novel |
|----------------|------------|-------------|-------------|-------------|
| k-means++ [1] | SODA/2007 | 77.1 | 70.1 | 80.7 |
| GCD [42] | CVPR/2022 | 80.2 | 85.1 | 77.6 |
| XCon [12] | BMVC/2022 | 86.7 | 91.5 | 84.1 |
| DCCL [30] | CVPR/2022 | 88.1 | 88.2 | 88.0 |
| SimGCD [52] | ICCV/2023 | 88.8 | 84.9 | 90.9 |
| InfoSieve [33] | NIPS/2023 | 91.8 | 92.6 | 91.3 |
| MOS (our) | | 93.2 | 89.5 | 95.1 |

we set the balancing factor $\lambda = 0.35$ and the temperature parameters $\tau_u = 0.07$ and $\tau_c = 1.0$, in line with the settings in [52]. The temperature values for classification losses are set to $\tau_t = 0.07$ and $\tau_s = 0.1$. We fine-tune the training hyperparameters on the FGVC-Aircraft. For FGVC-Aircraft, we apply a weight decay of $5e-4$ and perform a warm-up of τ_t from 0.04 to 0.07 after 20 epochs. All experiments run on a single NVIDIA GeForce RTX 4090 GPU.

4.4. Results

The results from Tab. 2 demonstrate that our MOS method achieves a significant leap forward in fine-grained image classification. Specifically, our approach outperforms the baseline SimGCD with an average improvement of **9%** on three parts of the Semantic Shift Benchmark (SSB). In comparison, our method also exhibits superior performance, achieving an average enhancement of **5%**. The results in-

Table 4. **Comparison of framework components.** **Object** refers to using object images for traditional training. **MOS** denotes the model excluding all components except the scene-awareness module. **SA Module** refers to the scene-awareness module itself. Each component contributes significantly to the overall performance improvement.

| Object | MOS | SA Module | All | Base | Novel |
|--------|-----|-----------|--------------|--------------|--------------|
| - | - | - | 61.50 | 65.70 | 59.40 |
| ✓ | - | - | 63.06 | 64.71 | 62.23 |
| ✓ | ✓ | - | 67.86 | 70.78 | 66.40 |
| ✓ | ✓ | ✓ | 69.57 | 72.31 | 68.20 |

indicate that scene information provides significant improvements in class inference for both base and novel classes.

We also assess **MOS** on Oxford-IIIT Pet dataset, ImageNet-100. The results clearly show that our approach significantly improves performance in comparison to previous methods. Specifically, our method outperforms the baseline SimGCD by **4.4%** in terms of accuracy.

4.5. Ablation Study

To better understand the performance improvements achieved by our proposed **MOS** framework, we conduct an ablation study on the CUB dataset, with detailed results shown in Tab. 4. We decompose the MOS framework into its individual components and evaluate each one separately to quantify their contributions. The results demonstrate that each component contributes positively to the overall performance, benefiting both base and novel classes.

We conduct a parameter ablation study, as shown in Fig. 5, to systematically assess the impact of different branch losses on model performance. The results demonstrate the effectiveness of the proposed balancing strategy and indicate that small deviations from the optimal configuration result in minimal performance degradation.

We explore the use of shared weights across multiple stages, as illustrated in Fig. 5. Results show that employing shared weights enhances both model performance, highlighting the advantages of this approach in improving both effectiveness and resource utilization. The ablation study on shared weights reveals that shared weights have a significant impact on the performance with novel classes, indicating that effective extraction of scene features plays a crucial role in improving the inference of novel classes.

4.6. Scene Information Analysis

We analyze whether the network has learned scene features by examining the differences between global image features and corresponding object features. Specifically, we track the mean deviation and L1 deviation of these differences throughout training. Mean deviation represents the central

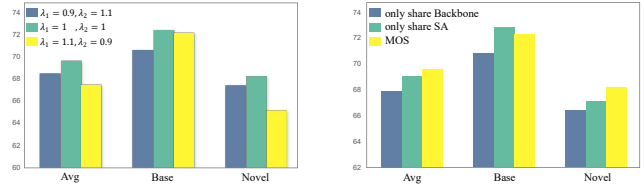


Figure 5. **Ablation Study on Parameters (left) and Shared Weights (right) on the CUB dataset.** The ablation study on parameters shows that the method is robust to parameter changes. The ablation study on shared weights reveals that shared weights primarily influence the inference of novel classes

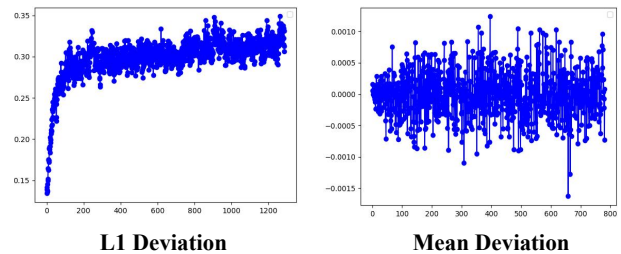


Figure 6. **Variation of mean deviation and L1 deviation between global image features and object features during training.** The x-axis represents the training iterations. The mean deviation (right) fluctuates within a small range (± 0.001), suggesting that the global image and object features form a unified and stable prototype. In contrast, the L1 deviation (left) increases notably, especially in the early stages of training, indicating the network’s growing ability to distinguish between the global image and object features as it learns to perceive scene differences.

deviation between the global image features and the object features. If the mean deviation fluctuates within a narrow range, it indicates that the feature prototypes of the global image and object are similar, suggesting that the network has formed stable and consistent class prototypes. L1 deviation measures the relative deviation between the global image features and the object features. Compared to object-centric feature extractors like DINO, we observe that the L1 deviation increases as training progresses. This indicates that the network is learning to distinguish the differences between the global image features and the object features. Together, these analyses show that the network is effectively learning to capture scene features, as evidenced by the growing divergence between global and object-specific features during training.

4.7. Visualization

t-SNE. The t-SNE visualization, showcased in Fig. 8, highlights the strengths of our method, showcasing significantly tighter clustering and more distinct category separation than both SimGCD and SPTNet. It proves that the incorporation of scene information effectively increases the inter-class margin.



Figure 7. **Segmentation Visualization on the CUB dataset.** It displays the segmentation results of IS-Net. Although the model produces several poor cases (Line 4), this does not undermine the overall effectiveness of our approach, highlighting its robustness against occasional failures.

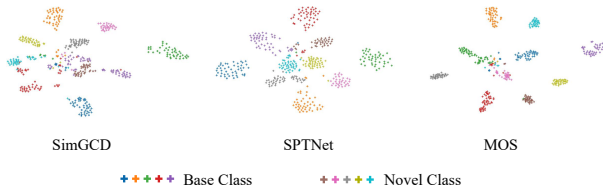


Figure 8. **The t-SNE visualization of representations of 10 classes randomly sampled from CUB dataset.** The results indicate that MOS is beneficial for maximizing inter-class margins.

Segmentation. Our current saliency segmentation model, IS-Net, is exclusively used for zero-shot segmentation inference, with no additional training. A common concern is whether our method is overly sensitive to segmentation performance fluctuations. We illustrate results of segmentation in Fig. 7. The saliency segmentation model produces many poor cases, yet our model still performs excellently, showcasing its robustness.

5. Conclusion

In this paper, we present a novel perspective on the role of scene information in Generalized Category Discovery (GCD), challenging the assumption that scenes act as noise in model training. Our findings demonstrate that scene information can serve as a crucial source of prior knowledge, enhancing the model’s ability to classify both base and novel categories. We identify the Ambiguity Challenge as a key factor contributing to the misinterpretation of

scene information in GCD tasks, where the overlap of objects from base and novel categories in different scenes can lead to misclassification. After addressing this challenge, we show that scene information can significantly improve classification performance. To leverage this insight, we propose the Modeling Object-Scene Associations (MOS) framework, which incorporates a simple MLP-based scene-awareness module that effectively distinguishes between scene and object features. Our experiments on fine-grained datasets demonstrate that MOS outperforms existing state-of-the-art methods. This result underscores the importance of integrating scene information into the GCD framework, rather than dismissing it as irrelevant.

Discussion. Our method achieves excellent performance in fine-grained datasets. However, it struggles to extract effective object information in extremely low-resolution settings, such as CIFAR 10/100.

6. Acknowledgment

This work is supported by the National Natural Science Foundation of China (62302167, U23A20343, 62472282, 72192821, 62302297), Shanghai Sailing Program (23YF1410500, 22YF1420300), Chenguang Program of Shanghai Education Development Foundation Shanghai Municipal Education Commission (23CGA34), and Young Elite Scientists Sponsorship Program by CAST (2022QNRC001).

References

- [1] David Arthur, Sergei Vassilvitskii, et al. k-means++: The advantages of careful seeding. In *Soda*, pages 1027–1035, 2007. 6
- [2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 3
- [3] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. *arXiv preprint arXiv:2102.03526*, 2021. 1, 3, 6
- [4] Xinzi Cao, Xiawu Zheng, Guan hong Wang, Weijiang Yu, Yunhang Shen, Ke Li, Yutong Lu, and Yonghong Tian. Solving the catastrophic forgetting problem in generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16880–16889, 2024. 6
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3
- [6] Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*, 2023. 3
- [7] Yanbei Chen, Xi Tian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3569–3576, 2020. 3
- [8] Yuhao Chen, Xin Tan, Borui Zhao, Zhaowei Chen, Renjie Song, Jiajun Liang, and Xuequan Lu. Boosting semi-supervised learning by exploiting all unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7548–7557, 2023. 3
- [9] Yujun Chen, Xin Tan, Zhizhong Zhang, Yanyun Qu, and Yuan Xie. Beyond the label itself: Latent labels enhance semi-supervised point cloud panoptic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1245–1253, 2024. 3
- [10] Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, and Ismail Ben Ayed. Parametric information maximization for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1729–1739, 2023. 1, 3, 6
- [11] Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive mean-shift learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23094–23104, 2024.
- [12] Yixin Fei, Zhongkai Zhao, Siwei Yang, and Bingchen Zhao. Xcon: Learning with experts for fine-grained category discovery. *arXiv preprint arXiv:2208.01898*, 2022. 1, 3, 6
- [13] Zhengyang Feng, Qianyu Zhou, Qiqi Gu, Xin Tan, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Dmt: Dynamic mutual training for semi-supervised learning. *Pattern Recognition*, 130:108777, 2022. 3
- [14] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9284–9292, 2021. 1, 3, 6
- [15] JY Gong, YJ Lou, FQ Liu, ZW Zhang, HM Chen, ZZ Zhang, X Tan, Y Xie, and LZ Ma. Scene point cloud understanding and reconstruction technologies in 3d space. *Journal of Image and Graphics*, 28(6):1741–1766, 2023. 3
- [16] Kai Han, Sylvestre-Alvise Rebuffi, Sebastian Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6767–6781, 2021. 1, 3, 6
- [17] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 5
- [18] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 3
- [19] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896, 2013. 3
- [20] Zechao Li, Jinhui Tang, and Tao Mei. Deep collaborative embedding for social image understanding. *IEEE transactions on pattern analysis and machine intelligence*, 41(9): 2070–2083, 2018. 3
- [21] Zechao Li, Yanpeng Sun, Liyan Zhang, and Jinhui Tang. Ctnet: Context-based tandem network for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9904–9917, 2021. 3
- [22] Zekun Li, Lei Qi, Yinghuan Shi, and Yang Gao. Iomatch: Simplifying open-set semi-supervised learning with joint inliers and outliers utilization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15870–15879, 2023. 3
- [23] Lizhuang Ma, Fei Wu, Qirong Mao, Pengjie Wang, and Yulong Chen. Visual recognition technologies for complex scenarios analysis. *Journal of Image and Graphics*, 2022. 3
- [24] Shijie Ma, Fei Zhu, Zhun Zhong, Xu-Yao Zhang, and Cheng-Lin Liu. Active generalized category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16890–16900, 2024. 3
- [25] Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5
- [26] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023. 3

- [27] Jona Otholt, Christoph Meinel, and Haojin Yang. Guided cluster aggregation: A hierarchical approach to generalized category discovery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2618–2627, 2024. 6
- [28] Rabah Ouldoung, Chia-Wen Kuo, and Zsolt Kira. Clipgcd: Simple language guided generalized category discovery. *arXiv preprint arXiv:2305.10420*, 2023. 1, 3
- [29] Zhengyuan Peng, Qijian Tian, Jianqing Xu, Yizhang Jin, Xuequan Lu, Xin Tan, Yuan Xie, and Lizhuang Ma. Generalized category discovery in semantic segmentation. *arXiv preprint arXiv:2311.11525*, 2023. 3
- [30] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptual contrastive learning for generalized category discovery. In *CVPR*, 2023. 1, 3, 6
- [31] Nan Pu, Wenjing Li, Xingyuan Ji, Yalan Qin, Nicu Sebe, and Zhun Zhong. Federated generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28741–28750, 2024. 3
- [32] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022. 3, 6
- [33] Sarah Rastegar, Hazel Doughty, and Cees Snoek. Learn to categorize or categorize to learn? self-coding for generalized category discovery. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3, 5, 6
- [34] Sarah Rastegar, Mohammadreza Salehi, Yuki M Asano, Hazel Doughty, and Cees GM Snoek. Selex: Self-expertise in fine-grained generalized category discovery. In *European Conference on Computer Vision*, pages 440–458. Springer, 2024. 3
- [35] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 3
- [36] Tianfang Sun, Zhizhong Zhang, Xin Tan, Yong Peng, Yanyun Qu, and Yuan Xie. Uni-to-multi modal knowledge distillation for bidirectional lidar-camera semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3
- [37] Tianfang Sun, Zhizhong Zhang, Xin Tan, Yanyun Qu, and Yuan Xie. Image understands point cloud: Weakly supervised 3d semantic segmentation via association learning. *IEEE Transactions on Image Processing*, 2024.
- [38] Shuai Tan, Biao Gong, Yutong Feng, Kecheng Zheng, Dandan Zheng, Shuwei Shi, Yujun Shen, Jingdong Chen, and Ming Yang. Mimir: Improving video diffusion models for precise text understanding. *arXiv preprint arXiv:2412.03085*, 2024.
- [39] Shuai Tan, Bin Ji, Mengxiao Bi, and Ye Pan. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *European Conference on Computer Vision*, pages 398–416. Springer, 2024.
- [40] Shuai Tan, Bin Ji, and Ye Pan. Flowvqtalker: High-quality emotional talking face generation through normalizing flow and quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26317–26327, 2024. 3
- [41] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3
- [42] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, 2022. 1, 3, 5, 6
- [43] Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3, 6
- [44] Andrea Vedaldi. Cats and dogs. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3498–3505, 2012. 5
- [45] Hongjun Wang, Sagar Vaze, and Kai Han. Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *International Conference on Learning Representations (ICLR)*, 2024. 1, 3, 6
- [46] Xuan Wang and Zhigang Zhu. Context understanding in computer vision: A survey. *Computer Vision and Image Understanding*, 229:103646, 2023. 3
- [47] Yu Wang, Pengchong Qiao, Chang Liu, Guoli Song, Xiawu Zheng, and Jie Chen. Out-of-distributed semantic pruning for robust semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23849–23858, 2023. 3
- [48] Yu Wang, Zhun Zhong, Pengchong Qiao, Xuxin Cheng, Xiawu Zheng, Chang Liu, Nicu Sebe, Rongrong Ji, and Jie Chen. Discover and align taxonomic context priors for open-world semi-supervised learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 3
- [49] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 5
- [50] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 3
- [51] Zhifeng Xie, Rui Qiu, Sen Wang, Xin Tan, Yuan Xie, and Lizhuang Ma. Pig: Prompt images guidance for night-time scene parsing. *IEEE Transactions on Image Processing*, 2024. 3
- [52] Bingchen Zhao Xin Wen and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *ICCV*, 2023. 1, 3, 4, 6
- [53] Yang Wang Yanan Wu, Zhixiang Chi and Songhe Feng. Metagcd: Learning to continually learn in generalized category discovery. In *ICCV*, 2023. 1, 3
- [54] Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*, 2024. 3
- [55] Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Khan. Promptcal: Contrastive affinity learning via auxiliary prompts

- for generalized novel category discovery. *arXiv preprint arXiv:2212.05590*, 2022. [1](#), [3](#), [6](#)
- [56] Bingchen Zhao and Oisín Mac Aodha. Incremental generalized category discovery. *arXiv preprint arXiv:2304.14310*, 2023. [3](#)
- [57] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. *arXiv preprint arXiv:2305.06144*, 2023.
- [58] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16623–16633, 2023. [1](#), [6](#)
- [59] Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. In *European Conference on Computer Vision*, pages 41–58. Springer, 2024. [3](#)