Semantic-Guided Novel Category Discovery

Weishuai Wang[1], Ting Lei[1], Qingchao Chen[2], Yang Liu [1]*

[1]Wangxuan Institute of Computer Technology, Peking University
[2]National Institute of Health Data Science, Peking University
wangweishuai@pku.edu.cn, ting_lei@pku.edu.cn, qingchao.chen@pku.edu.cn, yangliu@pku.edu.cn

Abstract

The Novel Category Discovery problem aims to cluster an unlabeled set with the help of a labeled set consisting of disjoint but related classes. However, existing models treat class names as discrete one-hot labels and ignore the semantic understanding of these classes. In this paper, we propose a new setting named Semantic-guided Novel Category Discovery (SNCD), which requires the model to not only cluster the unlabeled images but also semantically recognize these images based on a set of their class names. The first challenge we confront pertains to effectively leveraging the class names of unlabeled images, given the inherent gap between the visual and linguistic domains. To address this issue, we incorporate a semantic-aware recognition mechanism. This is achieved by constructing dynamic class-wise visual prototypes as well as a semantic similarity matrix that enables the projection of visual features into the semantic space. The second challenge originates from the granularity disparity between the classification and clustering tasks. To deal with this, we develop a semantic-aware clustering process to facilitate the exchange of knowledge between the two tasks. Through extensive experiments, we demonstrate the mutual benefits of the recognition and clustering tasks, which can be jointly optimized. Experimental results on multiple datasets confirm the effectiveness of our proposed method. Our code is available at https://github.com/wang-weishuai/Semantic-guided-NCD.

Introduction

Deep neural networks have surpassed human performance in various computer vision tasks. However, most traditional works focus on a closed-set setting, assuming training and testing data share the same class set, which results in limited generalization ability when the deep model is deployed in the wild. To address this limitation, Novel Category Discovery (NCD) (Hsu, Lv, and Kira 2018; Hsu et al. 2019; Han, Vedaldi, and Zisserman 2019; Han et al. 2020; Fini et al. 2021) attempts to train a network to cluster instances from unlabeled data, by utilizing labeled instances from a disjoint set of classes. The primary motivation behind this approach is to leverage the available supervision from the labeled set to learn powerful image representations that can be applied to cluster unlabeled instances.

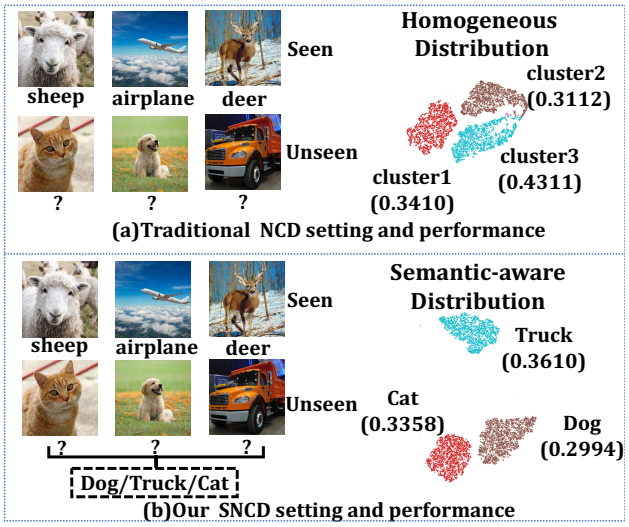


Figure 1: (a) Traditional NCD approaches perform class-agnostic clustering on unlabeled data, whose clustering results are homogeneous in distribution. (b) Our proposed approach, with a set of class names of unlabeled data as the only additive input, could get reasonable clustering results along with cluster categories. Moreover, the distribution of clusters is semantically related, meaning that clusters with closer semantic relations are more likely to be spatially proximate.

However, with the innate ability to comprehend the semantics of various types of instances, the human cognitive system excels not only in clustering instances that we have never seen before but also in associating them with previously known concepts (Tenenbaum et al. 2011). This remarkable capability empowers us to establish connections even when confronted with novel situations. When introduced to a new concept like a lynx, our brain might come up with other feline creatures such as domestic cats or larger relatives like tigers and lions. This innate ability to form associations with existing knowledge enables us to rapidly comprehend and recognize novel concepts like the lynx.

Motivated by this, we propose a novel Semantic-guided Novel Category Discovery (SNCD) setting, where an ad-

*Corresponding author

ditional set of class names for unlabeled images is available. This novel setting facilitates the exploration of semantic similarities across diverse classes and greatly assists in the clustering process for unlabeled categories. As shown in Fig. 1[1], by utilizing the set of names of unlabeled classes as the sole additional information, our model can acquire more reasonable semantic-aware clustering results rather than homogeneous one. Moreover, unlike previous NCD methods, our method gains the ability to semantically recognize each cluster by leveraging the pool of class names from unlabeled data as the supplementary information.

In the proposed SNCD setting, we encounter two major challenges that need to be addressed. The first challenge relates to the cross-modal gap between the visual and linguistic domains. Since the high-dimensional characteristics of visual features pose a hurdle in extending the visual space to the unlabeled classes, leveraging the natural generalizability of the linguistic domain has become a promising strategy. However, the intrinsic gap between two modalities makes it difficult to incorporate the semantic information of class names for recognizing unlabeled classes. The second challenge originates from the granularity difference between the classification and clustering task. Specifically, while the classification task requires sample-wise recognition based on the given class names, the clustering task merely requires to output a predefined number of clusters without knowing the semantics of unlabeled classes. As a result, it's non-trivial to optimize the two tasks simultaneously and make them complement each other. Besides, tackling the tasks of classification and clustering separately can result in sub-optimal outcomes.

To address the first challenge, we incorporate language priors derived from the linguistic domain along with visual-aware similarities, as shown in the classification branch of Fig. 2 shaded by blue. Specifically, we first acquire visual-aware classification scores through class-wise visual prototypes. Then, to leverage the relevance between different classes for better generalizability, we incorporate a semantic similarity matrix to facilitate the model to make *semantic-aware predictions* for both seen and unseen classes.

The second challenge entails developing a semantic-aware clustering process, where we incorporate linguistic knowledge from the classification branch into the clustering algorithm. As shown in Fig. 2, our objective is to merge the knowledge from the semantic-aware classification branch (shaded by blue) and the semantic-agnostic clustering branch (shaded by green) of our model to foster knowledge complementation in-between. To accomplish this, we first generate pseudo labels for the clusters predicted by the clustering branch, as presented in the Algorithm 1. Then we employ mutual information maximization between the classification score and clustering score to strengthen the collaboration between these two tasks.

[1]The number after every clustering is the variance of each cluster. The smaller this number is, the better the clustering performance is. Compared to the NCD approach, we could largely shrink the variance in every corresponding cluster, making the cluster more tight and trustworthy.

To summarize, our contributions can be outlined as follows: 1) We propose Semantic-guided Novel Category Discovery, a practical enhancement to the NCD setting that enables recognition of unlabeled data. 2) We design a dynamic class-wise visual prototype and a semantic similarity matrix to bridge the gap between visual and linguistic domains. 3) We propose a pseudo label generation method and mutual information maximization technique, which leverages the complementary knowledge from both the classification and clustering tasks, allowing for the concurrent optimization of both tasks. 4) Our approach achieves state-of-the-art performance on various existing benchmarks.

Related Work

Novel Category Discovery

Novel category discovery (NCD) aims to cluster instances in unlabeled data, by exploiting prior knowledge from known classes. NCD lies in transferring knowledge from labeled set to unlabeled set(Han et al. 2020, 2021a; Zhao and Han 2021; Zhong et al. 2021b). The task of NCD requires strong semantic similarity between labeled and unlabeled classes in order to group new instances. RS (Han et al. 2020, 2021a) generates pair-wise pseudo-labels by ranking-statistics. Zhao and Han (Zhao and Han 2021) further improves this method by utilizing local part-level information. NCL (Zhong et al. 2021a) and Jia (Jia et al. 2021) use contrastive learning to learn discriminative representations. UNO (Fini et al. 2021) proposes a unified training objective based on self-labeling. Joseph (Joseph et al. 2022) design a spacing loss to enforce separability in the latent space. ComEx (Yang et al. 2022) divides and conquers NCD with two groups of compositional experts. GCD (Vaze et al. 2022) proposes a more practical setting where unlabeled images may come from labeled classes or novel ones. However, all the methods and tasks outlined above treat image classes as discrete one-hot labels. In other words, semantic information of labeled and unlabeled categories, which could provide crucial inter-category cues for recognition or clustering, is overlooked in existing settings. Considering the above, we propose a new SNCD setup: given the set of class names for unlabeled images, the model is able to not only effectively cluster instances belonging to novel classes, but also assign a meaningful, semantic class name to every unlabeled image.

Zero-Shot Learning

Zero-shot learning (ZSL) aims to classify unlabeled classes out of the training set. In other words, ZSL aims to transfer the model from labeled to unlabeled classes. During training, the model is provided with objects or concepts along with their associated attributes. Recent works have achieved good performance on this task(Goodfellow et al. 2014; Arjovsky, Chintala, and Bottou 2017; Mishra et al. 2018; Chen et al. 2018; Schonfeld et al. 2019; Yu et al. 2020). However, the training process of ZSL is costly and requires a lot of semantic information as supervision, e.g., class attributes. Our classification branch can be seen as a simple but effective zero-shot model, which uses only class names we can easily

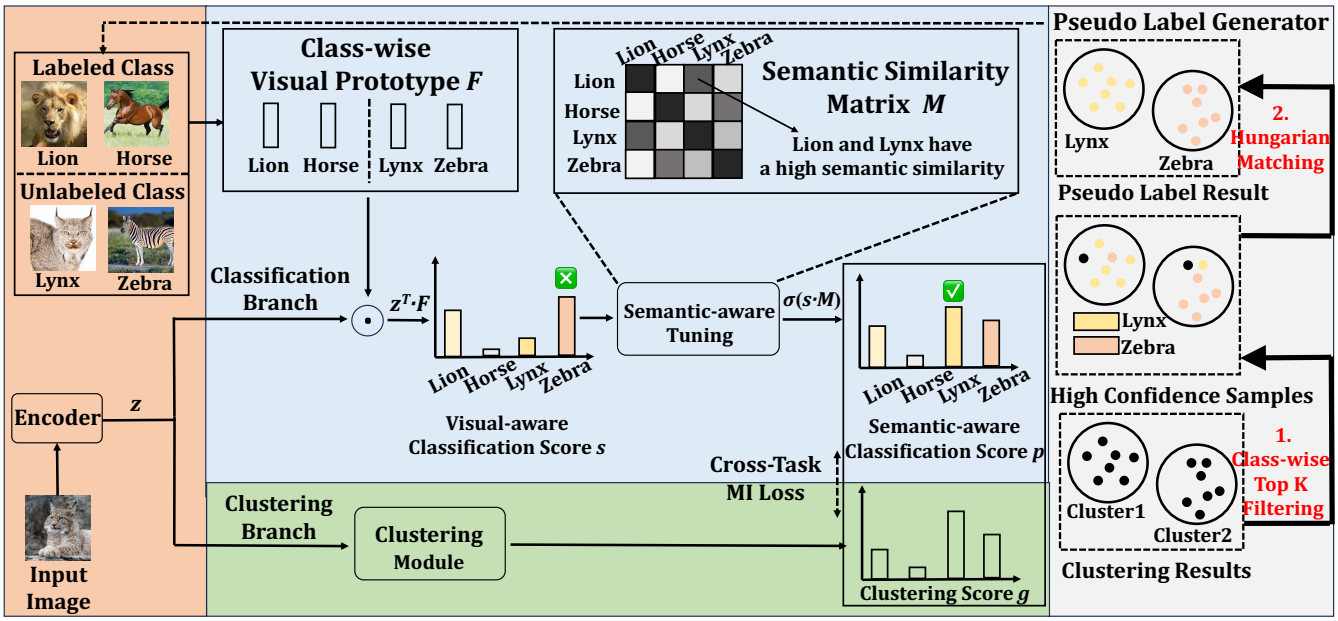


Figure 2: Overview of our approach for Semantic-guided Novel Category Discovery. In the classification branch, the visual feature z first engages with the class-specific Visual Prototype F, producing a visual-aware classification score s. The Semantic Similarity Matrix M is then employed to facilitate the infusion of semantic knowledge into s. Consequently, a refined semantic-aware classification score labeled as p is obtained on $(C_l + C_u)$ classes. The Visual Prototype is dynamically updated during training. In the clustering branch, we use (Zhang et al. 2022) based clustering module to get the cluster prediction g. To jointly optimize classification and clustering tasks, we propose to leverage the intra-cluster structure of the clustering results to generate reliable pseudo labels for unlabeled classes, providing high-quality candidates to update the class-wise visual prototype. Besides, we propose to maximize the mutual information of the predictions between the classification and clustering tasks.

get access to and achieves promising zero-shot classification results.

To highlight the uniqueness of our proposed new task SNCD, we summarize the differences in these tasks from the perspective of the information required during the training and testing phase. As shown in Figure 3, we can observe that SNCD, NCD, and ZSL tasks differ in terms of the information acquired during the training phase and the outputs during the testing phase. SNCD requires more information during the training phase compared to NCD and ZSL. However, during the testing phase, SNCD provides a combination of traditional NCD and ZSL outputs. NCD lacks the ability to recognize unlabeled images, while ZSL is not capable of clustering unlabeled images if the corresponding class names are not provided. SNCD can obtain both clustering and classification results, greatly enhancing the generalization ability of the task.

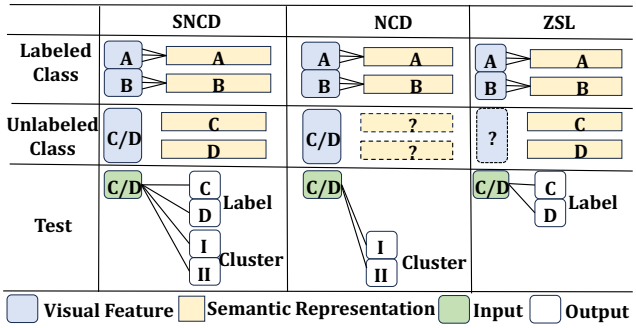


Figure 3: Comparison among different 'novel class understanding' task settings.

Method

As shown in Fig. 2, we adopt a dual-branch architecture, consisting of the classification branch F_C and the clustering branch F_G. The classification branch performs semantic-aware recognition, bridging the gap between the visual and linguistic domains. The clustering branch outputs the class-agnostic clustering results for the given images. We utilize mutual information between the outputs of two branches to facilitate the cooperation and complementation between them. In this section, we first introduce preliminaries, including the dataset setup and the clustering module. Then we discuss the incorporation of language priors with visual similarities to facilitate semantic-aware recognition. Besides, we discuss how the classification and clustering tasks interact with each other. Finally, we present the overall objective function of our method.

Preliminary

Problem Formulation. The training data is split into two different sets: a labeled set denoted as D_l =

$\{(x_1, y_1), ..., (x_N, y_N)\}$ and an unlabeled set denoted as $D_u = \{x_1, ..., x_M\}$. Each x_i represents an image, and each y_i signifies its respective label. We assume that there're C_l categories in D_l and C_u categories in D_u, and these categories are *disjoint but semantically connected*. Different from the NCD setting, in our SNCD setting, even though the individual labels are not known for the unlabeled data, we have prior knowledge of the set of possible class names for them.

Unlabeled Data Clustering. We adopt the clustering module in (Zhang et al. 2022)(denoted as F_G), which contains a visual encoder (shared with classification branch), a labeled head for labeled images classification, an unlabeled head for unlabeled images clustering and a mechanism to maximize mutual information between these two heads. During training phase, an labeled image with its class name or an unlabeled images is send to F_G, with an output $g = F_G(x)$ of $C_l + C_u$ dimension referring the probability assigning to each cluster, where x is the input image. Besides, we denote the overall loss for clustering module as $L_{cluster}$. More details can be found in (Zhang et al. 2022).

Semantic-Aware Classification

To effectively identify semantic similarities between labeled and unlabeled classes, we introduce a novel task that consists of not only clustering the unlabeled images but also classifying them. Our goal is to integrate language priors associated with visual features to improve generalization. Specifically, we propose a dynamic class-wise visual prototype and a semantic-aware tuning mechanism for this integration.

As depicted in figure 2, the proposed classification branch consists of two parts, the class-wise visual prototype F and Semantic Similarity Matrix M. Class-wise visual prototype F contains C_l features from the labeled set and C_u features from the unlabeled set. Given a visual feature z, we compute its cosine similarity with F to produce the visual classification score $s \in R^{C_l+C_u}$. Namely,

$$\boldsymbol{s} = \boldsymbol{z}^\top \cdot \boldsymbol{F}, \tag{1}$$

The mere visual similarities neglect the underlying relationships of semantics in linguistics between the labeled and unlabeled sets. Hence we further utilize the language priors embedded in M to enhance generalizability for unlabeled classes, where M_{ij} represents the cosine similarity between the word vectors of the i^{th} category and the j^{th} category. By supplementing these coefficients related to semantics, the final prediction of an unlabeled image will take into account its semantic similarity with the labeled class. Formally, the semantic-aware classification score $\boldsymbol{p}$ is calculated as:

$$\boldsymbol{p} = \sigma(\boldsymbol{s} \cdot \boldsymbol{M}), \tag{2}$$

where σ indicates the softmax function.

The semantic similarity matrix helps to revise some erroneous predictions in the visual-aware classification score s. For example, as illustrated in Fig. 2, lion and horse belong to labeled classes, while lynx and zebra belong to unlabeled classes. Consider a scenario where we aim to differentiate between a lynx and a zebra in an image. The prediction s_{zebra} erroneously surpasses s_{lynx}. To address this, we can leverage the visual-aware classification score and the semantic similarities among various classes as priors to revise the final prediction to $p_{zebra} < p_{lynx}$.

Dynamic Visual Prototype Maintenance. The Class-wise Visual Prototype F contains a single prototype for each class to capture the essential characteristics of each class. We randomly select K samples for a specific class and calculate their mean to represent the corresponding prototype. Enabling dynamic updates for the visual prototypes can significantly enhance the generalization performance of the model. Our update strategy differs for the labeled and unlabeled parts of the visual prototype.

For the labeled classes, the visual prototype is updated straightforwardly thanks to the availability of ground truth labels during the training process. Specifically, each training batch comprises images from various categories, including both labeled and unlabeled ones. The features z of the labeled images can be directly utilized to displace the corresponding old features in the visual prototype F based on their respective ground truth labels.

For the unlabeled classes, we employ a distinct update strategy. We first assign pseudo labels to all samples belonging to unlabeled classes. Then for each unseen class, we randomly select K samples from this pool and compute the mean feature of them. These mean features are then assigned to their respective slots within F. The details of pseudo labeling will be elaborated in the following subsection.

Coordination between the Classification and Clustering Tasks

As mentioned previously, handling classification and clustering tasks separately may lead to sub-optimal outcomes, such as homogeneous labeling results without reasoning, and unreliable classification results due to the lack of supervision for the unlabeled data. To mitigate these issues, we propose to make the two branches collaborate with each other, since the classification branch may offer noisy yet valuable information for unlabeled classes while the clustering branch excels at reliably dividing samples into clusters(Fini et al. 2021) without comprehending the semantics of each class. Specifically, we introduce two cooperative mechanisms to jointly optimize the two branches: 1) cluster-wise pseudo-label generation, and 2) mutual information maximization.

Cluster-Wise Pseudo-Label Generation: To provide features of unlabeled images with high confidence to visual prototype, which in turn leads to a better classification result, we introduce a cluster-wise pseudo-label generation process. The key objective of cluster-wise pseudo-label generation is to assign unique class labels to different clusters using the semantic-aware predictions from the classification branch. Specifically, in every training epoch, each image is first forwarded to both the *clustering and classification branches*, generating its predicted category name (based on $\boldsymbol{p}$) and clustering ID (based on $\boldsymbol{g}$), respectively. To leverage the semantic-aware classification results in the *classification branch*, we propose to select anchor images with *top-K confidence* for each unlabeled class, as illustrated in the process

of Category-wise Top K Filtering in Fig. 2.

The cluster-wise Pseudo-label Generation procedure is presented in Algorithm 1. Line 2 initializes a matrix $S \in R^{C\times C}$, where $S_{i,j}$ aims to record the number of high-quality samples belonging to the j^{th} class assigned to i^{th} cluster. In lines 3-4, we obtain the clustering and classification results by forwarding N images through the clustering and classification branches, respectively. In lines 5-7, we iterate through all the classes and find top-K samples for each class to form the matrix G_{temp}. As illustrated in the part of Pseudo Label Generator of figure 2, these anchor images may lie in different clusters from *the clustering branch*, especially at the beginning of training. The different numbers of class-wise anchor images *in each cluster* can be regarded as a pattern to assign a specific class for each cluster, as shown in lines 8-12 of the Algorithm 1. Then, leveraging these class-wise anchor patterns in multiple clusters, our goal is to find the permutation of the rows of S to maximize the trace of it. To achieve this, we utilize the Hungarian algorithm to match C_u clusters with C_u unlabeled categories. The resulting output Q consists of matched pairs, each containing a cluster id and its corresponding pseudo-label. Finally, we employ the matching results to associate the cluster id in G with the pseudo-label. The intuition is that the more high-confidence anchor images from a *class* are selected in a cluster, the more confident it is to assign the cluster to that class.

Till now, we could obtain high-quality cluster-wise pseudo labels by utilizing both the intra-cluster structure of clustering results and the semantic-aware confidence scores. We randomly select samples with pseudo labels and employ them to update the prototype for each unseen class, as detailed in the preceding subsection.

Mutual Information Maximization: To strengthen the collaboration and facilitate knowledge transfer between the classification and clustering tasks, we propose to maximize the mutual information between the class prediction probabilities obtained from the clustering branch and those obtained from the classification branch:

$$\boldsymbol{P} = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{g}(\boldsymbol{x}_b)\cdot \boldsymbol{p}(\boldsymbol{x}_b)^T \tag{3}$$

$$I(\boldsymbol{g},\boldsymbol{p}) = \sum_{i=1}^{C_l+C_u}\sum_{j=1}^{C_l+C_u} \boldsymbol{P}_{ij}\log\frac{\boldsymbol{P}_{ij}}{\boldsymbol{p}_i\boldsymbol{p}_j^{zsl}} \tag{4}$$

where B is the batch size, $\boldsymbol{g}$ is the outputs of the clustering branch, $\boldsymbol{p}$ is the predicted probabilities in the classification branch and P represents the joint probability distribution of the output logits of clustering and classification branches. $L_t^{MI} = -I(g,\boldsymbol{p}))$ is minimized to maximize the mutual information between the classification and clustering tasks.

The classification branch may produce unreliable classification results due to the lack of supervision for the unlabeled data. However, compared with the predictions from the classification branch, the clustering branch provides a more reliable understanding regard to class relationships. Hence we maximize mutual information between the output of the clustering branch and the classification branch to merge the knowledge of the two branches.

Algorithm 1: Cluster-wise Pseudo-label Generation

Input: N Image features $I \in \mathbb{R}^{N\times d}$, a hyperparameter K

1: **for** epoch in range(TotalEpochs) **do**
2: S = zeros(C,C)
3: G=$F_G(I)\in \mathbb{R}^{N\times C}$ {forward via clustering Branch}
4: P=$F_C(I)\in \mathbb{R}^{N\times C}$ {forward via classification Branch}
5: **for** i in range(C) **do** {check every category}
6: indice = argmax Top$K$ P[:,i] {Top K Image indices most likely belonging to the category i}
7: G_{temp} = G[indice,:] {$G_{temp} \in \mathbb{R}^{K\times C}$}
8: **for** j in range(K) **do**
9: clusterID = argmax G_{temp}[j,:]
10: S[clusterID,i] += 1
11: **end for**
12: **end for**
13: Q = Hungarian-Matching(S) {find the matching between the cluster id and pseudo label}
14: $PseudoLabel$ = Label-Assignment(G, Q) {utilize the matching results to generate pseudo labels for the clustering results}
15: **end for**

Overall Objective

During training, our model jointly performs the classification and clustering tasks on labeled and unlabeled data. The network is optimized by the following objective:

$$L = L^{cluster} + \alpha L_t^{MI} \tag{5}$$

where α is a hyper-parameter to balance the loss terms. During inference, we use $argmax(\boldsymbol{g})$ to specify the cluster id, and $argmax(\boldsymbol{p})$ to infer the class for unlabeled data.

Experiments

This section presents a comprehensive evaluation of our method on three NCD benchmarks. We focus on both clustering and classification accuracy to showcase our method's superiority over existing methods. We also conduct ablation study to explore the impact of individual modules.

Experimental Setup

Datasets. We evaluate our method on three benchmark NCD datasets: CIFAR10, CIFAR100 (Krizhevsky et al. 2009), and ImageNet (Deng et al. 2009). We follow the dataset splits of various settings in (Fini et al. 2021).

Metrics. We employ two primary metrics to evaluate our method's performance. For classification tasks, we utilize classification accuracy as the metric. For the clustering task, we adhere to the metric introduced in (Fini et al. 2021).

Protocol. We assess our method under both **task-aware** and **task-agnostic** circumstances. In the task-aware protocol, only novel(unlabeled) categories are evaluated, and it is known whether an image belongs to a labeled or unlabeled class. The task-agnostic protocol, proposed by Fini (Fini et al. 2021), simulates real-world conditions where distinguishing between labeled and unlabeled categories is nec-

Method	CIFAR10		CIFAR100-20		CIFAR100-50		ImageNet	
	Classification	Clusering	Classification	Clusering	Classification	Clusering	Classification	Clusering
k-means	-	72.5	-	56.3	-	28.3	-	71.9
KCL	-	72.3	-	42.1	-	-	-	73.8
MCL	-	70.9	-	21.5	-	-	-	74.4
DTC	-	88.7	-	67.3	-	35.9	-	78.3
RS	-	90.4	-	73.2	-	39.2	-	82.5
RS+	-	91.7	-	75.2	-	44.1	-	82.5
UNO	-	**96.1**	-	84.5	-	52.8	-	89.2
UNOv2	-	93.6	-	90.2	-	61.0	-	91.1
Ours	**40.1**	93.8	**57.8**	**93.7**	**21.6**	**62.2**	**26.5**	**92.5**

Table 1: Comparison with state-of-the-art methods on CIFAR-10, CIFAR-100, and ImageNet on classification and clustering metrics, using task-aware evaluation protocol. '-' means the methods treat class names as discrete one-hot labels, lacking semantic understanding of classes and therefore cannot perform the classification task.

Method	CIFAR10			CIFAR100-20			CIFAR100-50		
	Labeled	Unlabeled	All	Labeled	Unlabeled	All	Labeled	Unlabeled	All
KCL	79.4	60.1	69.8	23.4	29.4	24.6	-	-	-
MCL	81.4	64.8	73.1	18.2	18.0	18.2	-	-	-
DTC	58.7	78.6	68.7	47.6	49.1	47.9	30.2	34.7	32.5
RS+	90.6	88.8	89.7	71.2	56.8	68.3	69.7	40.9	55.3
UNO	93.5	**93.3**	93.4	73.2	72.7	73.1	71.5	50.6	61.0
UNOv2	95.3	91.3	93.3	73.6	71.0	73.1	73.4	57.5	65.5
Ours	**95.8**	92.7	**94.3**	**79.9**	**79.2**	**79.5**	**77.2**	**60.5**	**68.9**

Table 2: Comparison with state-of-the-art methods on CIFAR-10 and CIFAR-100 on both labeled and unlabeled classes, using task-agnostic evaluation protocol.

Method	CIFAR100-20
DEM (Zhang, Xiang, and Gong 2017)	19.8
f-CLSWGAN (Xian et al. 2018)	27.3
FREE (Chen et al. 2021)	35.6
CE-GZSL (Han et al. 2021b)	37.5
Ours	**57.8**

Table 3: Comparative results under ZSL settings.

essary. We can not know whether an image comes from labeled set or unlabeled set in advance at test time.

Implementation Details. Our experiments are built on the UNO baseline (Fini et al. 2021), keeping the hyperparameters consistent. We reproduce the performance metrics reported in the original papers if available. The encoder f_θ uses a ResNet18 (He et al. 2016) pretrained on the labeled data set for the classification task. The weight α for the mutual information loss is set to 0.1. We choose $K = 16$ when we generate pseudo labels for unlabeled classes and use the Glove word vectors (Pennington, Socher, and Manning 2014) trained by Wikipedia2014 and Gigaword5 to supply semantic information.

Comparison with Other Methods

This part compares state-of-the-art models and our model's ability to deal with the SNCD task. We compare our method with k-means (MacQueen et al. 1967), KCL (Hsu, Lv, and Kira 2018), MCL (Hsu et al. 2019), DTC (Han, Vedaldi, and Zisserman 2019), RS, RS+ (Han et al. 2020), UNO (Fini et al. 2021), UNOv2. As shown in Table 1, our model achieves more significant performance gains on 80/20 split than 50/50 split of CIFAR100, showing that given more concepts/semantics, our model can better mine semantic similarities between labeled classes and unlabeled classes. Concretely, for the CIFAR100-20 split, our method achieves nearly a 4.5% increase in clustering accuracy, while for ImageNet, our approach surpasses the leading NCD method by 1%. Besides, unlike previous methods that treat classes as discrete one-hot labels, our model is semantic-aware and can perform classification task, achieving a promising zero-shot classification result.

For the task-agnostic setting, Table 2 reveals the significant improvements our method brings to both labeled and unlabeled sets. On the unlabeled set of CIFAR100-20, our approach demonstrates a 9% performance boost when compared to the leading method. Additionally, we achieve a performance gain of 3% to 4% on both CIFAR100-20 and CIFAR100-50 splits.

Furthermore, we compare the classification performance of unlabeled classes between traditional zero-shot image recognition methods and our semantic-guided classification branch using the CIFAR100-20 split. To ensure fairness, the

zero-shot learning methods are implemented using the conventional ZSL setting, employing the same feature encoder and word vectors as our framework. Our method surpasses all the ZSL baselines mentioned, as demonstrated in Table 3. This highlights the efficacy of our approach and showcases its proficiency in uncovering semantic similarities between known and unknown classes while demonstrating a profound comprehension of unlabeled classes.

Ablation Study

Network Component: To validate the effectiveness of our proposed components, we conduct an ablation study focusing on both classification and clustering accuracy on the CIFAR100-20 split, as shown in Table 4. We observe: (1) Note that our model without $F^{unlabel}$ has to incorporate the semantic similarity matrix M to classify images belonging to unseen classes. Comparing the results presented in lines 1 and 2, we can find an 18.3% and a 1.4% improvement in classification and clustering accuracy, respectively. This proves that mutual information between the classification and clustering tasks leverages the underlying data correlations to enhance the overall performance. (2) When comparing the results presented in lines 2 and 4, it becomes evident that incorporating unlabeled class samples ($F^{unlabel}$) leads to a 5.5% enhancement in classification performance. This is because some semantic information of the unlabeled classes cannot be well captured solely through the representation of labeled classes. By enabling the integration of image features originating from the unlabeled classes into the class-wise visual prototype F, a more comprehensive representation of these unlabeled images, along with their associated semantic labels, can be achieved. (3) By comparing the outcomes observed in lines 3 and 4, we find that projecting the visual similarity s into the semantic space through the Semantic Similarity Matrix M yields a noteworthy enhancement, which proves that the capacity for generalization within the semantic space surpasses that of the visual space.

F^{label}	$F^{unlabel}$	M	L_t^{MI}	CIFAR100-20	
				Classification	Clustering
✓	✗	✓	✗	34.0	92.4
✓	✗	✓	✓	52.3	93.0
✓	✓	✗	✓	50.8	92.7
✓	✓	✓	✓	**57.8**	**93.7**

Table 4: Ablation study on the network component.

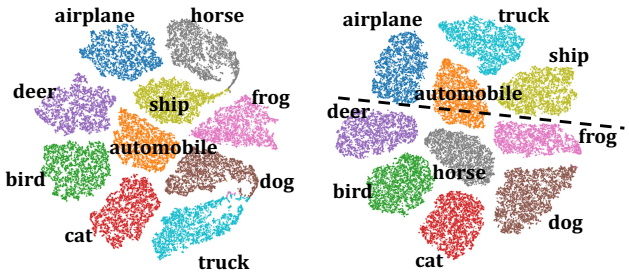
Implementation of Cross-Task Knowledge Transfer: We conduct ablation experiments on different loss functions to enhance the effectiveness of cross-task knowledge transfer. As shown in Table 5, we find that using L2-distance to supervise the correlation between the output of two tasks leads to inferior performance. This is because clustering score represents the cluster id, which does not contain information related to the class name. In contrast, mutual information is capable of quantifying the relationship between two random variables that are sampled concurrently, making it a more powerful approach in the context of joint optimization. It enables a soft knowledge transfer mechanism that ensures consistency without strictly enforcing identical predictions.

Method	CIFAR100-20	
	Classification	Clustering
Mutual Information	57.8	93.7
L2-distance	30.2	88.6

Table 5: Ablation on the implementation of cross-task knowledge transfer.

Visualization Analysis

To evaluate the ability of our proposed clustering module to learn from the recognition task, we visualize the logits $\boldsymbol{g}$ for both UNO and our method, as shown in figure 4. In the case of UNO, it erroneously identifies cats and dogs as the closest neighbors of trucks, showing a homogeneous distribution. This result is unreasonable as it fails to consider the semantic similarity between these categories. However, our method effectively addresses this issue by appropriately clustering trucks closer to other vehicle categories such as airplanes, cars, and ships. This pattern demonstrates the effectiveness of our approach in grouping images based on their semantic similarities, resulting in more logical and accurate clustering outcomes.



(a) UNO **(b) our SNCD**

Figure 4: t-SNE visualization for all classes on CIFAR10 for UNO and our method.

Conclusion

In this paper, we propose the task of Semantic-guided Novel Category Discovery. Our objective is to develop a model that can effectively identify semantic similarities among various classes given a set of class names about unlabeled images. We design a dynamic visual prototype and a semantic-aware tuning strategy for interaction between visual features and semantic labels. In addition, we develop a semantic-aware clustering process to transfer knowledge between the classification and clustering branches. We also show that the clustering and classification modules can improve each other's performance. Extensive experiments on three widely used datasets prove the effectiveness of our proposed method.

Acknowledgements

This work was supported by the grants from the National Natural Science Foundation of China 62372014.

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Chen, L.; Zhang, H.; Xiao, J.; Liu, W.; and Chang, S.-F. 2018. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1043–1052.

Chen, S.; Wang, W.; Xia, B.; Peng, Q.; You, X.; Zheng, F.; and Shao, L. 2021. Free: Feature refinement for generalized zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 122–131.

Deng, J.; Socher, R.; Fei-Fei, L.; Dong, W.; Li, K.; and Li, L.-J. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*.

Fini, E.; Sangineto, E.; Lathuilière, S.; Zhong, Z.; Nabi, M.; and Ricci, E. 2021. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9284–9292.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*.

Han, K.; Rebuffi, S.; Ehrhardt, S.; Vedaldi, A.; and Zisserman, A. 2021a. AutoNovel: Automatically Discovering and Learning Novel Visual Categories. *TPAMI*.

Han, K.; Rebuffi, S.-A.; Ehrhardt, S.; Vedaldi, A.; and Zisserman, A. 2020. Automatically Discovering and Learning New Visual Categories with Ranking Statistics. In *Proc. ICLR*.

Han, K.; Vedaldi, A.; and Zisserman, A. 2019. Learning to discover novel visual categories via deep transfer clustering. In *Proc. ICCV*.

Han, Z.; Fu, Z.; Chen, S.; and Yang, J. 2021b. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2371–2381.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hsu, Y.-C.; Lv, Z.; and Kira, Z. 2018. Learning to cluster in order to transfer across domains and tasks. In *Proc. ICLR*.

Hsu, Y.-C.; Lv, Z.; Schlosser, J.; Odom, P.; and Kira, Z. 2019. Multi-class classification without multi-class labels. In *Proc. ICLR*.

Jia, X.; Han, K.; Zhu, Y.; and Green, B. 2021. Joint Representation Learning and Novel Category Discovery on Single- and Multi-modal Data. In *Proc. ICCV*.

Joseph, K.; Paul, S.; Aggarwal, G.; Biswas, S.; Rai, P.; Han, K.; and Balasubramanian, V. N. 2022. Spacing Loss for Discovering Novel Categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3761–3766.

Krizhevsky, A.; et al. 2009. Learning multiple layers of features from tiny images. *University of Tronto*.

MacQueen, J.; et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proc. BSMSP*.

Mishra, A.; Krishna Reddy, S.; Mittal, A.; and Murthy, H. A. 2018. A generative model for zero shot learning using conditional variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2188–2196.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8247–8255.

Tenenbaum, J. B.; Kemp, C.; Griffiths, T. L.; and Goodman, N. D. 2011. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, 331(6022): 1279–1285.

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2022. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7492–7501.

Xian, Y.; Lorenz, T.; Schiele, B.; and Akata, Z. 2018. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5542–5551.

Yang, M.; Zhu, Y.; Yu, J.; Wu, A.; and Deng, C. 2022. Divide and Conquer: Compositional Experts for Generalized Novel Class Discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14268–14277.

Yu, Y.; Ji, Z.; Han, J.; and Zhang, Z. 2020. Episode-based prototype generating network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14035–14044.

Zhang, C.; Hu, C.; Xu, R.; Gao, Z.; He, Q.; and He, X. 2022. Mutual Information-guided Knowledge Transfer for Novel Class Discovery.

Zhang, L.; Xiang, T.; and Gong, S. 2017. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021–2030.

Zhao, B.; and Han, K. 2021. Novel Visual Category Discovery with Dual Ranking Statistics and Mutual Knowledge Distillation. *arXiv preprint arXiv:2107.03358*.

Zhong, Z.; Fini, E.; Roy, S.; Luo, Z.; Ricci, E.; and Sebe, N. 2021a. Neighborhood Contrastive Learning for Novel Class Discovery. In *Proc. CVPR*.

Zhong, Z.; Zhu, L.; Luo, Z.; Li, S.; Yang, Y.; and Sebe, N. 2021b. OpenMix: Reviving Known Knowledge for Discovering Novel Visual Categories in An Open World. In *Proc. CVPR*.