Novel Class Discovery for Ultra-Fine-Grained Visual Categorization

Yu Liu[1], Yaqi Cai[1], Qi Jia[1]*, Binglin Qiu[1], Weimin Wang[1], Nan Pu[2]
[1]International School of Information Science and Engineering, Dalian University of Technology
[2]Department of Information Engineering and Computer Science, University of Trento
{liuyu8824,jiaqi,wangweimin}@dlut.edu.cn, {caiyaqi1998,m1andy}@mail.dlut.edu.cn,
nan.pu@unitn.it

Abstract

Ultra-fine-grained visual categorization (Ultra-FGVC) aims at distinguishing highly similar sub-categories within fine-grained objects, such as different soybean cultivars. Compared to traditional fine-grained visual categorization, Ultra-FGVC encounters more hurdles due to the small inter-class and large intra-class variation. Given these challenges, relying on human annotation for Ultra-FGVC is impractical. To this end, our work introduces a novel task termed Ultra-Fine-Grained Novel Class Discovery (UFG-NCD), which leverages partially annotated data to identify new categories of unlabeled images for Ultra-FGVC. To tackle this problem, we devise a Region-Aligned Proxy Learning (RAPL) framework, which comprises a Channel-wise Region Alignment (CRA) module and a Semi-Supervised Proxy Learning (SemiPL) strategy. The CRA module is designed to extract and utilize discriminative features from local regions, facilitating knowledge transfer from labeled to unlabeled classes. Furthermore, SemiPL strengthens representation learning and knowledge transfer with proxy-guided supervised learning and proxy-guided contrastive learning. Such techniques leverage class distribution information in the embedding space, improving the mining of subtle differences between labeled and unlabeled ultra-fine-grained classes. Extensive experiments demonstrate that RAPL significantly outperforms baselines across various datasets, indicating its effectiveness in handling the challenges of UFG-NCD. Code is available at https://github.com/SSDUT-Caiyq/UFG-NCD.

1. Introduction

"There are no two identical leaves in the world", Leibniz once said. However, it is trivial and troublesome to identify the subtle differences given visually similar leaf images, even for senior botanists. Traditional fine-grained visual categorization (FGVC) [1] has enabled the differentiation of subordinate categories within a broader class, such as bird species [26], by focusing on distinct region features like head and wing. Yet, FGVC methods are still limited when it comes to dissecting these categories further into sub-categories. To this end, ultra-fine-grained visual categorization (Ultra-FGVC) [20, 27, 30–34] is researched at recognizing sub-categories of ultra-fine-grained (UFG) images from a specific fine-grained category, which attracts the attention of researchers due to its significant potential in real-world applications like precision agriculture. In Fig. 1, for example, each row of leaf images corresponds to a distinct soybean cultivar, but the differences between these UFG classes are hardly able to be defined and perceived

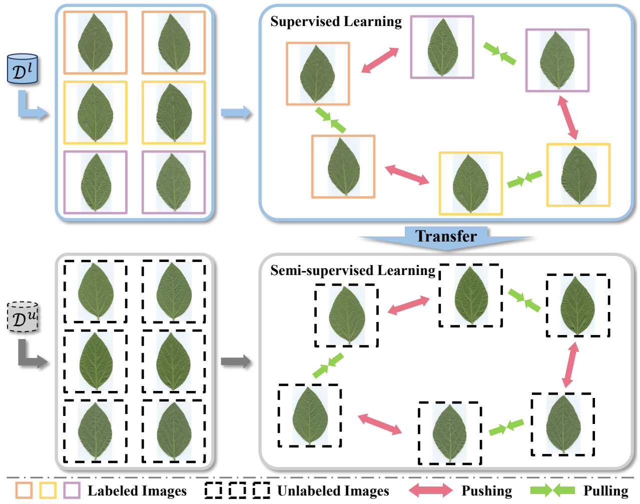


Figure 1. Visual conception of UFG-NCD, where D^l, D^u are labeled and unlabeled data splits. We propose to exploit novel class discovery for partially annotated UFG images. The core idea is to learn prior knowledge from labeled images via supervised learning, subsequently extending the knowledge to unlabeled images via a semi-supervised framework. It facilitates knowledge transfer and representation learning across ultra-fine-grained classes.

*corresponding author

by humans [31]. One can expect that, Ultra-FGVC suffers more from small inter-class and large intra-class variation inherent to FGVC. Hence, it is impractical to fully annotate UFG image instances for a supervised learning paradigm.

In recent years, an emergent task termed novel class discovery (NCD) [7–10, 16, 28] gathers more attention due to its ability to discover novel categories from a pool of unlabeled instances in a semi-supervised fashion. Note that, NCD task has access to some labeled classes disjoint from the unlabeled ones. The success of NCD relies on conceptual knowledge transfer from labeled to unlabeled images. It means that the model pre-trained on labeled data needs to learn discriminative representation suitable for clustering unlabeled samples. A line of early works is dedicated to modeling the relation of unlabeled images via clustering objectives [9] or soft class predictions [7, 10]. However, these approaches neglect the semantic relations between labeled and unlabeled classes, thereby hindering the transfer learning of discriminative representation. To this end, recent works [8, 16] focus more on capturing instance-level constraints in the feature or prediction space. Despite the empirical effectiveness of such methods on generic and fine-grained NCD benchmarks, they are limited by learning imperceptible clues from unlabeled UFG images.

In this work, we exploit novel class discovery for addressing Ultra-FGVC, with no need to fully annotate all the images. This new task setting termed UFG-NCD has not been studied in the research community yet. To tackle this problem, we propose a simple yet effective framework called region-aligned proxy learning (RAPL), which is tailored specifically for UFG-NCD. First of all, we develop a region-aligned module on the feature maps, which groups different channels and forces each group to discover discriminative features in a local region. In addition, inspired by proxy-based methods [2, 13, 39] in deep metric learning, we present a novel semi-supervised proxy learning (SemiPL) framework, modeling global structure in feature embedding space and facilitating representation learning by using the relation between class-specified proxies and data samples. In this way, the labeled and unlabeled data can be optimized jointly by proxy-guided supervised learning (PSL) and proxy-guided contrastive learning (PCL), respectively. Last but not least, our RAPL is also applicable to generalized category discovery (GCD) setting [21, 25], where the unlabeled data include both labeled and unlabeled categories. Our contributions are summarized as follows:

- We are the first to explore how to categorize unlabeled UFG images with labeled ones of disjoint classes, resulting in a new and practical task termed as UFG-NCD.
- We propose a channel-wise region alignment module to reveal subtle differences within a local region, which promotes the learning and transfer of local representation.
- We introduce a novel two-stage semi-supervised proxy learning framework for UFG-NCD, which models the relationship between instances and class proxies in embedding space, and learns from unlabeled images with global distribution information.
- We introduce five UFG-NCD datasets, where extensive experiments demonstrate our method achieves significant improvements over previous NCD methods.

2. Related Work

2.1. Ultra-Fine-Grained Visual Categorization

Yu et al. [31] pioneered the formulation of the ultra-fine-grained visual categorization (Ultra-FGVC) task, with the objective of discerning sub-categories within specific fine-grained categories. Concurrently, they introduced two ultra-fine-grained (UFG) datasets focused on leaves, namely SoyAgeing and SoyGene. The collected images of the same sub-category correspond to specific seeds obtained from the genetic resource repository. To address the challenges posed by the small inter-class and large intra-class variations in Ultra-FGVC, existing approaches typically fall into three categories: local-based [30, 32], feature fusion-based [20, 27, 34], and contrastive learning-based [33] methods. First, akin to conventional fine-grained classification [3, 29, 35], local-based approaches focus on learning local embeddings by predicting masked or erased regions within images. For example, MaskCOV [30] employed a covariance matrix to capture mutual information between each quarter of randomly masked and shuffled image patches. Second, feature fusion-based approaches enhance local and global discrimination by fusing important tokens, drawing inspiration from the attention mechanism in vision transformers. For instance, FFVT [27] introduced a novel mutual attention weight selection module that chose the token most similar to each anchor token. Third, motivated by the success of contrastive learning in self-supervised learning, CLE-ViT [33] generated positive samples through randomly masked and shuffled image parts, thereby mitigating the challenges of class variation. Notably, the effectiveness of these approaches relies on fully labeled data, which is challenging and expensive to obtain for UFG images in practical scenarios.

In contrast to these methods, our work enables the categorization of unlabeled UFG images in a semi-supervised learning fashion, leveraging partially labeled data.

2.2. Novel Class Discovery

Different from conventional semi-supervised learning [15, 19, 24], novel class discovery (NCD) infers unlabeled images from related but disjoint classes with labeled ones. In order to take advantage of labeled data, most of the approaches [7–10, 12, 16, 17, 28, 36–38] conduct two training stages. In the first stage, the model is pre-trained with the la-

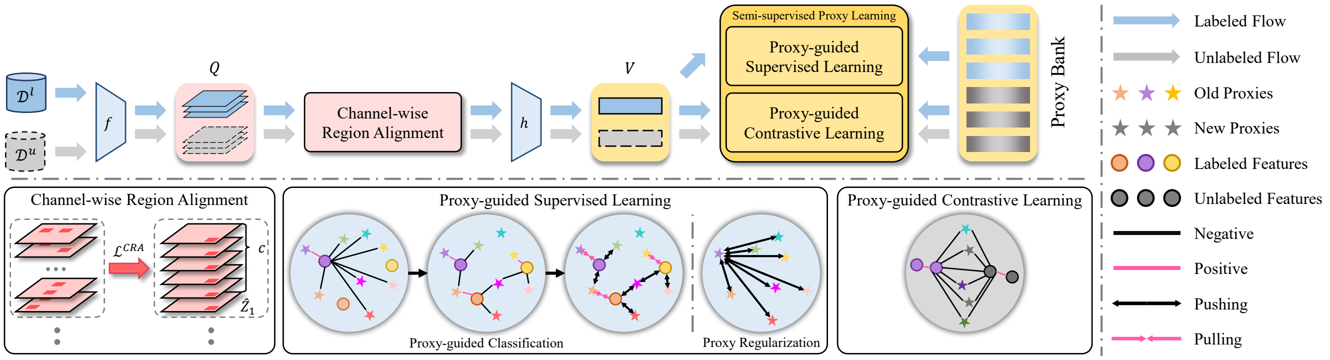

$\mathcal{D}^l$
$\mathcal{D}^u$
f
Q
Channel-wise Region Alignment
h
V
Semi-supervised Proxy Learning
Proxy-guided Supervised Learning
Proxy-guided Contrastive Learning
Proxy Bank
Labeled Flow
Unlabeled Flow
Old Proxies
New Proxies
Labeled Features
Unlabeled Features
Negative
Positive
Pushing
Pulling
Channel-wise Region Alignment
$\mathcal{L}^{CRA}$
c
$\hat{Z}_1$
Proxy-guided Supervised Learning
Proxy-guided Classification
Proxy Regularization
Proxy-guided Contrastive Learning

Figure 2. Overview of our Region-Aligned Proxy Learning (RAPL) framework. We first group and align extracted feature maps in the set Q with regions via Channel-wise Region Alignment. Then, we generate global representation v through global average pooling and MLP projection h, and lastly learn and transfer knowledge with the proxy bank using semi-supervised proxy learning (SemiPL).

beled data for a standard supervised learning objective; for the second stage, the knowledge learned from the model is transferred to group the unlabeled data based on elaborately designed clustering or self-supervised objectives. Specifically, early work [10, 36] modeled the binary relationship between samples based on the similarity of the top-k feature dimensions. UNO [7] built a connection of images and classes by the Sinkhorn-Knopp algorithm [4]. Meanwhile, a line of work [37, 38] employed richer inter-sample relations with nearest neighbor in feature space. However, these instance-level or feature-level methods neglect rich inter-class relations, leading to compact class distributions that hinder the representation learning of new categories. Thus, recent works [8, 16] were dedicated to leveraging inter-class similarity information. OpenLDN [22] further explored pairwise similarity relations to regularize feature distributions of labeled and unlabeled data. Nevertheless, these methods are much challenged by small inter-class variance in UFG sub-categories. In this work, we proposed a novel semi-supervised proxy learning framework to explicitly learn inter-class discriminative knowledge through class-specified proxies in feature space.

3. Proposed Method

3.1. Problem Formulation

Denote $\mathcal{D}$ as a UFG dataset, including a labeled split $\mathcal{D}^l = \{(x_i^l, y_i^l)\}_{i=1}^{N^l}$ and an unlabeled split $\mathcal{D}^u = \{(x_i^u, y_i^u)\}_{i=1}^{N^u}$, where y is the class label of image x, and N^l and N^u are the number of instances in two data splits. It is noteworthy that, following NCD, we assume labeled images from old classes, *i.e.*, $y^l \in \mathbb{R}^{C^l}$, while unlabeled ones are from disjoint new classes to be classified, *i.e.*, $y^u \in \mathbb{R}^{C^u}$, and $C^l \cap C^u = \emptyset$. During the training phase, we assume labels y^l are available, whereas y^u remain unknown. Our objective is to classify the unlabeled split based on the labeled one. Consistent with most prior research [7, 10], the number of new classes C^u is assumed to be a known priori.

3.2. Overview

As depicted in Fig. 2, we propose a novel region-aligned proxy learning (RAPL) framework that consists of three main designs for UFG-NCD. First of all, we develop a novel channel-wise region alignment (CRA) module, which aims at explicitly learning local fine-grained representation in distinct regions with their corresponding groups of feature maps. After locating regional patterns via CRA, we introduce a semi-supervised proxy learning (SemiPL) approach, integrating proxy-guided supervised learning (PSL) and proxy-guided contrastive learning (PCL) simultaneously. PSL learns a global representation and regularizes the distribution of embedding space via proxy-guided classification and proxy regularization. Besides, we adopt PCL to facilitate representation learning on unlabeled data with global distribution information of class proxies.

3.3. Channel-wise Region Alignment

The distinctions among ultra-fine-grained classes become less perceptible compared to classes in conventional FGVC. Consequently, categorizing these classes heavily depends on local discriminative features. For instance, MaskCOV [30] and SPARE [32] attend local information via semantic context constraints across randomly masked or erased image patches. These methods mainly model patch-to-patch relations in a certain image for local representation learning, whereas such relations are unable to be shared and transferred from labeled samples to unlabeled ones. Thus, we propose a channel-wise region alignment (CRA) module to leverage local representation for distinctive feature learning in UFG-NCD. As illustrated in the bottom-left subplot of Fig. 2, we explicitly organize and associate feature channels with specific regions to extract subtle visual cues within each region. The local features learned in CRA are discriminative for all classes, advancing the transfer of knowledge from labeled classes to unlabeled ones.

Specifically, given an input image x_i, the feature maps extracted by an image encoder f is denoted as $Z_i = \{z_i^1, .., z_i^D\} \in \mathbb{R}^{D \times W \times H}$. Here, Z_i consists of D channels of feature matrix z with dimensions of height H and width W, where each dimension of z encodes a region representation of the image. In order to assign distinct weights to each local region, we divide D feature channels into $H * W$ groups, denoted as $\{\hat{Z}_i^0, \dots, \hat{Z}_i^{H*W-1}\}$, Each group is responsible for encoding a specific region representation. In this way, each group of feature channels is defined as

$$\hat{Z}_i^j = \{z_i^k \in \mathbb{R}^{H \times W} | k \in [j \times c + 1, .., j \times c + c]\}, \quad (1)$$

where each group comprises $c = \frac{D}{H*W}$ channels. Subsequently, a region label $y_i^k = j$ is assigned to each channel z_i^k in the group $\hat{Z}_i^j$, explicitly linking feature channels within each group to a specific region. As a result, we enforce each feature channel to concentrate on its corresponding region through a loss function denoted as $\mathcal{L}^{CRA}$, calculated by applying standard cross-entropy on each flattened feature matrix, *i.e.*, flatten$(z_i) \in \mathbb{R}^{H*W}$:

$$\mathcal{L}^{CRA} = \frac{1}{D} \sum_{k=1}^{D} -y_i^k \log(\text{flatten}(z_i^k)). \quad (2)$$

After learning regional patterns via CRA, it is crucial to combine them into a global representation for the final classification. As shown in Fig. 2, we achieve a global representation via global average pooling and MLP projection on the flattened feature maps, which are denoted as set of feature vectors $V = h(\text{flatten}(Q)) \in \mathbb{R}^{N \times D}$, where $h(\cdot) = \text{MLP}(\text{GAP}(\cdot))$, $Q = \{Z_1, ..Z_N\}$.

3.4. Semi-supervised Proxy Learning

In the context of Ultra-FGVC, the primary challenge is to capture subtle differences among similar samples that prove challenging to discern in feature space. The use of a linear classifier is a common approach to optimizing feature representation. However, this method is susceptible to overfitting on labeled classes, leading to the encoding of insufficient and less discriminative knowledge when faced with unlabeled classes [25]. Thus, we propose a novel semi-supervised proxy learning (SemiPL) approach for UFG-NCD, which encodes discrimination and distribution of labeled instances via proxy-guided supervised learning, and further transfers suitable knowledge from labeled images to unlabeled ones by proxy-guided contrastive learning.

Proxy-guided Supervised Learning (PSL). Rather than assigning a specific pseudo-label to each instance, we introduce class-level proxies to establish relations between instances and classes, and further learn representation and regularize distribution in feature space via proxy-guided classification and proxy regularization, respectively, drawing inspiration from works such as [2, 13, 39].

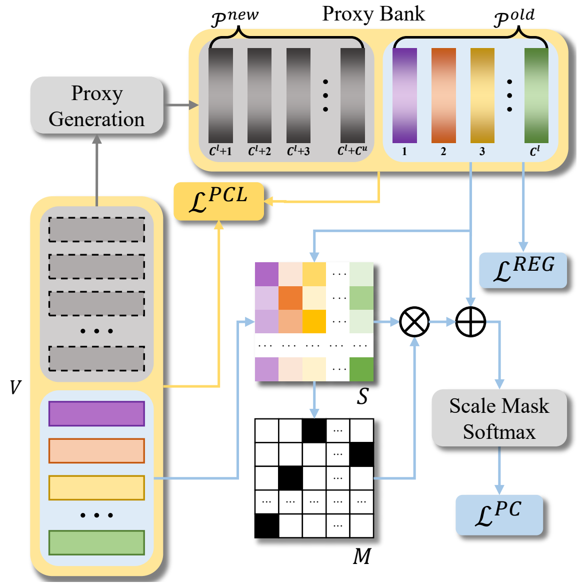


Figure 3. Illustration of proposed SemiPL. We first introduce a supervised paradigm, indicated by blue lines, which includes $\mathcal{L}^{PC}$ and $\mathcal{L}^{REG}$, representing the loss of proxy-guided classification and proxy regularization, respectively. Then, we propose PCL with global distribution guidance of $\mathcal{P}^{old}$ and $\mathcal{P}^{new}$ to learn and discover new classes from unlabeled data, indicated in yellow lines.

For proxy-guided classification, we first generate classification probabilities by comparing each visual feature with its $\{k+1\}$-nearest proxies, instead of learning a linear classifier. We aim to lower the probabilities of negative proxies by loss constraints during training. Initially, we construct a set of learnable proxies $\mathcal{P}^{old}$ according to labeled classes in feature space, which is denoted as $\mathcal{P}^{old} = \{(p_i^l, y_i^l) | i \in [1, .., C^l], p_i^l \in \mathbb{R}^D, y_i^l = i\}$. Subsequently, given features $V^l = \{v_1^l, .., v_{N^l}^l\}$ of labeled images, we compute the classification probabilities with cosine similarity:

$$S = \{s_{ij} = (v_i^l)^{\mathrm{T}} p_j^l | i \in [1, .., N^l], j \in [1, .., C^l]\}, \quad (3)$$

where s_{ij} indicates the similarity between the i-th instance v_i^l and the j-th proxy p_j^l. In the leftmost circle of the proxy-guided classification subplot in Fig. 2, let us take the feature labeled by the purple circle as an example, and the colorful stars represent the proxies. The pink line indicates the only positive relation between the feature and the proxy to which it belongs, while the black lines represent the negative relations between the feature and other proxies.

In conventional classification scenarios, s_{ij} should be significantly higher than others if the i-th instance belongs to the j-th class proxy. However, for Ultra-FGVC, each instance may have several negative relations with high probabilities across different classes due to small intra-class variation. To lower the high probabilities in negative relations at training phase, we select the top k negative relations using a mask matrix M with the same size as S. In particular, the selected probabilities in M are marked with 1, while others are marked with 0. Formally, each element in the mask $M \in \mathbb{R}^{N^l \times C^l}$ is defined by:

$$m_{ij} = \begin{cases} 1, & \text{if } s_{ij} \in \text{top-}k(S_i \backslash s_{ig}),\ g = y_i \\ 0, & \text{else}, \end{cases} \tag{4}$$

where $S_i = [s_{i1}, .., s_{iC^l}]$ represents the similarity between the i-th instance and all old proxies, and s_{ig} is the similarity between i-th instance and proxy of the correct class y_i to which the instance belongs, which is the probability of positive relation. Then, the class prediction improved by the mask becomes

$$\overline{Y} = S \odot M + S^{pos}, \tag{5}$$

where $S^{pos} = \{s_{ig} \in S | g = y_i\}$ is the set only including probabilities of positive relations. As shown in the second circle of the middle bottom subfigure in Fig. 2, each feature has one positive relation and k negative relations with corresponding proxies. Note that, $\overline{Y}$ is sparse due to a limited k, which leads to an inflated denominator in conventional softmax. Thus, we employ an indicator function $\mathbb{1}(\cdot)$ to eliminate the effect of redundant zero elements in $\overline{Y}$, and thereby the final prediction is

$$\tilde{Y}_{ij} = \frac{\mathbb{1}(\overline{Y}_{ij} \neq 0)\text{exp}(\overline{Y}_{ij})}{\sum_{g=1}^{C^l} \mathbb{1}(\overline{Y}_{ig} \neq 0)\text{exp}(\overline{Y}_{ig})}, \tag{6}$$

where $\mathbb{1}(\cdot)$ is 1 when $\overline{Y}_{ij} \neq 0$; otherwise is 0. Based on the prediction probabilities, we can define the loss of proxy-guided classification as

$$\mathcal{L}^{PC} = -\frac{1}{N^l} \sum_{i=1}^{N^l} \sum_{j=1}^{C^l} \mathbb{1}(y_i = j)\text{log}(\tilde{Y}_{ij}). \tag{7}$$

As shown in the third circle at the bottom of Fig. 2, the positive relation between the feature and corresponding proxy are pulled close while negative relations are pushed away during training.

As each proxy represents the center of a class, we expect the difference between proxies to be as large as possible. Furthermore, we introduce a regularization constraint to make the learned proxies as discriminative as possible. We compute the similarity between old proxies by

$$S^p = \{s^p_{ij} = (p^l_i)^{\mathrm{T}} p^l_j | i, j \in [1, .., C^l]\}. \tag{8}$$

Afterward, the proxy regularization is defined as

$$\mathcal{L}^{REG} = -\frac{1}{C^l} \sum_{i=1}^{C^l} \sum_{j=1}^{C^l} \mathbb{1}(i = j)\text{log}(\text{softmax}(S^p_{ij})). \tag{9}$$

It is worth noting that $\mathcal{L}^{REG}$ is conducted in feature space, which regularizes class distribution via similarity between class-specified proxies. The fourth circle at the bottom of Fig. 2 illustrates the proxies are pulled away properly thanks to imposing proxy regularization.

Proxy-guided Contrastive Learning (PCL). Considering the knowledge transfer and learning of unlabeled data, it is difficult to define the positive relations between unlabeled instances. Like recent work [7, 37], we take advantage of contrastive learning to learn instance-wise representation with two augmented views of each sample.

Before developing proxy-guided contrastive learning (PCL), we firstly capture global structure information of unlabeled classes by generating a set of proxies in feature space accordingly, which is marked as $\mathcal{P}^{new} = \{(p^u_i, y^u_i) | i \in [C^l + 1, .., C^l + C^u], p^u_i \in \mathbb{R}^D, y^u_i = i\}$. As shown in yellow lines of Fig. 3, we employ the guidance of both $\mathcal{P}^{old}$ and $\mathcal{P}^{new}$ to mine discriminative features in contrastive learning. Thus, the input vectors $\overline{V}$ of contrastive learning become a combination of v, v', and $\mathcal{P}^{old}$, $\mathcal{P}^{new}$, where the v, v' are the feature vectors extracted from two views of images. Note that, the positive vector for each proxy is itself in constrastive learning. The loss of PCL is written as

$$\mathcal{L}^{PCL} = -\sum_{i=1}^{|\overline{V}|} \text{log} \frac{\text{exp}(\overline{v}_i \cdot \overline{v}'_i / \tau)}{\sum_{j=1}^{|\overline{V}|} \mathbb{1}(i \neq j)\text{exp}(\overline{v}_i \cdot \overline{v}_j \tau)}, \tag{10}$$

where $\overline{v}_i, \overline{v}'_i$ is a feature vector and its positive vector in contrastive learning, τ is the temperature, and $|\overline{V}|$ is number of feature vectors in $\overline{V}$. Proxy-guided contrastive loss learns instance-level representation based on region-aligned feature maps and separate unlabeled data points in the embedding space. Note that, PCL achieves knowledge transfer without introducing any inaccurate positive relation, which optimizes feature distribution with the guidance of proxies.

3.5. Optimization Objective

Our training process comprises a pre-train phase and a discover phase. In the pre-train phase, we train the whole network with the proposed CRA and PSL modules on $\mathcal{D}^l$ to learn discriminative region representation and generic distribution information in embedding space:

$$\mathcal{L}_{pre-train} = \alpha \mathcal{L}^{CRA} + \beta \mathcal{L}^{PC} + \gamma \mathcal{L}^{REG}, \tag{11}$$

where weight parameters α, β and γ control each loss term. Specifically, $\mathcal{L}^{CRA}$ encourages the network to explore discriminative local features, $\mathcal{L}^{PC}$ pull the features and their corresponding proxies close, and $\mathcal{L}^{REG}$ promotes regularized class distribution in embedding space.

In the discover phase, we initialize proxy set $\mathcal{P}^{new}$ by k-means [18] on unlabeled data in $\mathcal{D}^u$. Note that, $\mathcal{P}^{old}$ are

Method	SoyAgeing-R1			SoyAgeing-R3			SoyAgeing-R4			SoyAgeing-R5			SoyAgeing-R6		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
RankStats IL [10]	42.12	71.92	12.32	42.53	71.31	13.74	40.10	69.70	10.51	40.71	68.69	12.73	33.33	52.93	13.74
UNO [7]	45.90	59.80	32.00	47.46	62.22	32.69	46.47	57.98	34.95	46.93	58.79	35.07	38.75	45.25	32.24
ComEx [28]	48.59	64.85	32.32	47.48	64.04	30.91	46.59	58.99	34.19	47.73	60.81	34.65	41.37	50.71	32.02
IIC [16]	53.24	71.91	34.55	51.57	70.51	32.63	47.73	60.81	34.65	49.60	61.82	37.37	40.20	49.09	31.31
rKD [8]	52.33	72.12	32.53	53.74	75.15	32.32	50.81	68.48	33.13	52.63	71.92	33.33	46.26	62.42	30.10
RAPL (Ours)	**58.99**	**79.19**	**38.79**	**58.99**	**78.59**	**39.39**	**57.07**	**71.92**	**42.22**	**61.01**	**74.75**	**47.27**	**50.20**	**64.65**	**35.76**

Table 1. Compared results on test splits of SoyAgeing-{R1, R3, R4, R5, R6} with task-agnostic protocol. Rankstats IL represents the implementation of an incremental classifier. The best results are marked in bold, and the second-best results are underlined.

Method	R1	R3	R4	R5	R6
RankStats IL [10]	12.12	14.55	11.72	11.92	10.71
UNO [7]	31.92	33.37	34.30	34.91	31.07
ComEx [28]	33.13	32.32	33.94	35.15	33.33
IIC [16]	34.34	34.55	35.71	34.44	32.07
rKD [8]	33.74	32.73	35.35	34.14	32.53
RAPL (Ours)	**43.03**	**46.06**	**49.90**	**45.25**	**41.82**

Table 2. Compared results on unlabeled train splits of SoyAgeing-{R1, R3, R4, R5, R6} with task-aware protocol.

fixed during this phase for providing consolidated guidance to distinguish old classes and recognize new classes. Thus, we removed proxy regularization in $\mathcal{L}_{pretrain}$ and incorporate proxy-guided contrastive learning in discover phase:

$$\mathcal{L}_{discover} = \alpha\mathcal{L}^{CRA} + \beta\mathcal{L}^{PC} + \delta\mathcal{L}^{PCL}, \qquad (12)$$

where δ is weight parameter for proxy-guided contrastive learning. For the final label assignment, we perform k-means directly on the extracted feature vector V.

4. Experiment

4.1. Experimental Setup

Data. Our experiments choose five datasets from [31] for UFG-NCD task, namely SoyAgeing-{R1, R3, R4, R5, R6}. These datasets contain leaves from five sequential growth stages of soybeans. Following the protocol in [10], we take the initial 99 categories as C^l, and the subsequent ones as C^u. For NCD scenario, each class encompasses 5 images for both training and testing. For generalized class discovery (GCD), there are 8 training images and 2 testing images per class, where unlabeled training data consists of 50% of the images from C^l, in addition to all images from C^u.

Evaluation Protocol. Following [10, 25], we evaluate baselines with classification accuracy for labeled images and the average clustering accuracy (ACC) for unlabeled ones. Given the application of k-means [18] for assigning classes to all images, ACC calculations are extended across all data. The ACC is defined as $ACC = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(y_i = perm_{max}(\hat{y}_i))$, where N is the number of instances for clustering, and $perm_{max}$ is the optimal permutation derived from Hungarian algorithm [14]. Following [7, 8, 10], we evaluate all methods with *task-agnostic* and *task-aware* protocols, respectively, indicating separate or joint assessments of instances from both old and new classes.

Implementation Details. To obtain more discriminative representation, following common implementation in Ultra-FGVC [30, 32], we train all methods using a ResNet-50 [11] backbone with ImageNet [5] pre-trained weights, where the dimension $D = 2048$ for feature maps and feature vector. Furthermore, akin to [25], we project the feature maps through a three-layer MLP projection head before applying SemiPL, and discard it at test-time. We resize input images into 448×448 and perform standard data augmentation as [33], generating two views for each image in the training process. For our RAPL, we optimize the whole architecture using SGD with momentum [23]. For each training phase, the initial learning rate is 0.01 which we decay with a linear warm-up and cosine annealed scheduler , and weight decay is set to 10^{-4}. Specifically, we first pre-train the network on D^l for 100 epochs, and then jointly fine-tune the network on both D^l and D^u for another 100 epochs. The batch size is set to 24, temperature τ is set to 0.1 in contrastive loss, and β is set to 2 and 1 in the pre-train and discover phase respectively. We adjust the loss weights based on ACC of "All" classes, as detailed in Sec. 4.4. All experiments are conducted on a single RTX 4090 GPU. More comprehensive details are in the Appendix.

4.2. Main Results

We adopt five representative NCD methods acting as baselines of UFG-NCD, including RankStats IL[10], UNO [7], ComEX [28], IIC [16], and rKD [8]. Results for all methods, using both task-agnostic and task-aware protocols, are detailed in Tab. 1 (test subset) and Tab. 2 (unlabeled training subset), respectively. The accuracy results in terms of "All", "Old", and "New", which denote all instances, and instances from old or new classes.

Overall, as shown in Tab. 1, our RAPL consistently outperforms all baselines by a significant margin. Specifically,

	Component					SoyAgeing-R1 [31]		
	$\mathcal{L}^{PC}$	$\mathcal{L}^{REG}$	$\mathcal{L}^{CRA}$	$\mathcal{L}^{PCL}$	$\mathcal{L}^{VCL}$	All	Old	New
(1)	✓	✓	✓		✓	56.36	74.14	38.59
(2)	✓	✓		✓		56.97	75.56	38.38
(3)	✓		✓	✓		58.48	76.97	**40.00**
(4)	✓	✓	✓	✓		**58.99**	**79.19**	38.79

Table 3. Ablation study. Each component of our method is removed in isolation, where $\mathcal{L}^{VCL}$ indicates vanilla contrastive learning without the guidance of proxies.

our method outperforms the best results of baselines, which from state-of-the-art NCD methods, *i.e.*, IIC[16] or rKD[8] with a minimum gain of 3.94% (on SoyAgeing-R6), and a maximum of 8.38% (on SoyAgeing-R5) for "All" classes. Interestingly, while rKD shows superior accuracy for "Old" classes, it underperforms in learning "New" classes, likely due to its explicit knowledge distillation of old classes to prevent catastrophic forgetting. Our RAPL method demonstrates consistent improvements in both "Old" and "New" categories, proving the effectiveness of our CRA strategy in capturing generic local representation for UFG images. This success is further bolstered by knowledge learning and transfer via SemiPL, which integrates classification knowledge within class-specified proxies and enhances feature-proxy similarity relations.

The results under task-aware protocol in Tab. 2 show that our method outperforms baselines on unlabeled training images. In particular, our method gains more considerable improvements than in Tab. 1, *e.g.*, 8.49% on SoyAgeing-R6, 14.19% on SoyAgeing-R4. Note that, SoyAgeing-R6 represents the final growth cycle of the soybean, where leaves of each class differentiate into various appearances which leads to large intra-class variation and makes representation learning harder. Thus, the results of SoyAgeing-R6 are relatively lower than those on other UFG datasets, because insufficient representation learning on labeled data leads to weak knowledge transfer on unlabeled ones.

4.3. Ablation Study

In Tab. 3, we report the ablation results on SoyAgeing-R1 by individually removing each component of RAPL, *i.e.*, $\mathcal{L}^{PC}$, $\mathcal{L}^{REG}$, $\mathcal{L}^{CRA}$ and $\mathcal{L}^{PCL}$, represents proxy-guided classification, proxy regularization, channel-wise region alignment, and proxy augmented contrastive learning.

Contrastive Learning. We verify the effectiveness of PCL by performing vanilla contrastive learning without the guidance of proxies. Comparing the Rows (1) and (4), PCL outperforms vanilla contrastive learning by 2.63% on "All" classes for SoyAgeing-R1. This implies that with the guidance of global distribution information of proxies in embedding space, the network learns a more comprehensive representation of UFG images in the discover stage.

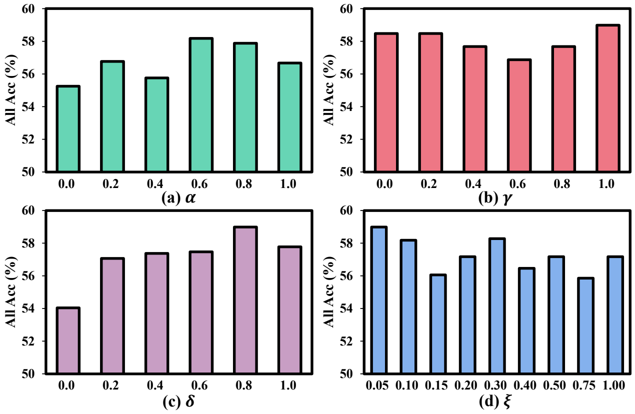


Figure 4. Impact of hyper-parameters on ACC with respect to "All" instances in the test dataset of SoyAgeing-R1.

Channel-wise Region Alignment. Rows (2) and (4) show the effect of the channel-wise region alignment for UFG-NCD. We observe a near 2% increase across ACC on "All" classes, with a more pronounced increase for the "old" class. This demonstrates that our CRA serves as a platform for effective representation learning and knowledge transfer. Especially when encountering labeled data, label information encourages learning discriminative features within specific common regions via $\mathcal{L}^{CRA}$. Besides, $\mathcal{L}^{CRA}$ facilitates representation learning on unlabeled data by sharing local knowledge across data samples and proxies.

Proxy Regularization. We observe a further performance improvement across ACC on "All" and "Old" classes after introducing proxy regularization from Rows (3) and (4). This indicates that regularization on $\mathcal{P}^{old}$ further learns the difference between each class which is represented by proxies and embeds global distribution information with $\mathcal{P}^{old}$, which is conducive to separating old classes on embedding space with relaxed margins.

4.4. Hyper-Parameter Analysis

In this section, we investigate the impact of the hyper-parameters in each training phase.

Impact of Loss Weights. We study the effect of each loss weight when fixing others. Specifically, we first choose the optimal α when γ and δ are both set to 1.0, then we determine the best γ based on selected α and select the optimal δ similarly. The impact of different values of each loss weight is shown in Fig. 4(a)-(c). Eventually, the hyper-parameters used in our method are set to $\alpha = 0.6, \gamma = 1.0$, and $\delta = 0.8$.

Impact of Number of Selected Proxies in $\mathcal{L}^{PC}$. We study the effect of k with above optimal loss weights, where we define $k = \xi \cdot C^l$ in $\mathcal{L}^{PC}$. As shown in Fig. 4(d), the best performance is achieved when $\xi = 0.05$, indicating that focus on separating similar classes is conducive to mining the subtle differences between UFG images.

Method	SoyAgeing-R1			SoyAgeing-R3			SoyAgeing-R4			SoyAgeing-R5			SoyAgeing-R6		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
GCD [10]	39.65	49.24	34.85	38.64	47.73	34.09	39.90	45.71	36.99	40.15	50.00	35.23	37.79	46.72	33.33
XCon [6]	33.84	42.93	29.29	33.08	42.68	28.28	35.52	40.15	33.21	39.98	49.24	35.35	34.09	41.92	30.18
RAPL	**52.36**	**60.89**	**48.08**	**50.87**	**58.47**	**47.07**	**53.30**	**62.10**	**48.89**	**51.14**	**61.69**	**45.86**	**44.01**	**52.42**	**39.80**

Table 4. Compared result on train split of SoyAgeing-{R1, R3, R4, R5, R6} for GCD task.

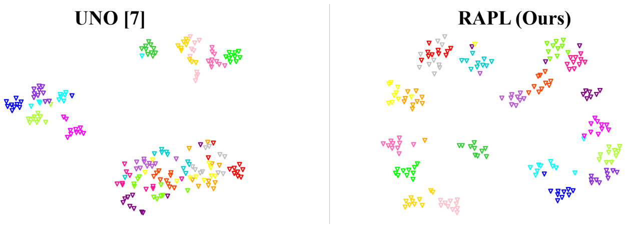
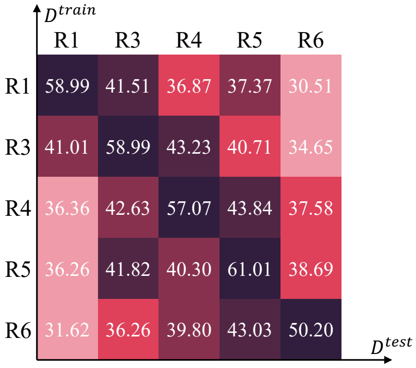


Figure 5. Visualization of features distributions of 20 unlabeled classes from SoyAgeing-R5.



	R1	R3	R4	R5	R6
R1	58.99	41.51	36.87	37.37	30.51
R3	41.01	58.99	43.23	40.71	34.65
R4	36.36	42.63	57.07	43.84	37.58
R5	36.26	41.82	40.30	61.01	38.69
R6	31.62	36.26	39.80	43.03	50.20

Figure 6. Generalization performance with RAPL, where y-axis indicates the train set and x-axis is the test set.

4.5. Further Investigation of RAPL

Evaluation on GCD Setting. To further evaluate the effectiveness of RAPL, we implement RAPL and two representative algorithms [6, 25] on the UFG dataset under the GCD setting. Note that, we use ResNet-50 with ImageNet pretrained weights for GCD and XCon for a fair comparison. As shown in Tab. 4, our RAPL consistently outperforms the GCD methods by over 10% on SoyAgeing-{R1,R3,R4,R5}, 6.22% on SoyAgeing-R6. Since classification heads must be trained from scratch, they are vulnerable to overfitting on the labeled classes after the pre-train phase of NCD. To this end, GCD and XCon turn to optimize the network with contrastive learning for more generalized feature representation in one stage. Nevertheless, the results reproted on Tab. 4 demonstrate that optimizing the network with RAPL boosts both representation learning on labeled and unlabeled instances, and avoids overfitting on the old data.

Visualization of Feature Distributions. To qualitatively explore the clustered features on the UFG dataset, we visualize the t-SNE embeddings based on UNO [7] and our RAPL. Clearly, the feature embeddings optimized with our RAPL become more discriminative and less overlap in the embedding space than UNO.

Generalization Performance of RAPL. Since SoyAgeing-{R1, R3, R4, R5, R6} are collected in different growth stages of soybean, where larger numbers represent later stages, we further investigate the generalization capability of RAPL when testing on the different datasets. As shown in Fig. 6, RAPL achieves greater generalization capability between nearly growth stages, *e.g.*, the performance of RAPL when training with SoyAgeing-R6 decreases on other datasets as the growth stage decreases. This tendency indicates that we can explore knowledge transfer across images of different ultra-fine-grained datasets to promote the classification of unlabeled UFG images.

5. Conclusion

We have proposed a task exploiting novel class discovery for ultra-fine-grained visual categorization, termed as UFG-NCD. To tackle this problem, we propose a simple yet effective framework named region-aligned proxy learning (RAPL), promoting the representation learning of local regions and the knowledge transfer from labeled to unlabeled images. Our RAPL is comprised of a channel-wise region alignment module to extract local representation, and a novel semi-supervised proxy learning framework to facilitate representation learning based on the relations between class-level proxies and instances. Experimental results on five datasets show that RAPL achieves state-of-the-art performance. In the future, RAPL has the potential to be applied to other ultra-fine-grained scenarios.

Acknowledgement. This work was supported by the NSF of China under Grant Nos. 62102061, 62272083 and 62306059, and by the Liaoning Provincial NSF under Grant 2022-MS-137 and 2022-MS-128. Besides, this work was partially supported by the EU Horizon projects ELIAS (No. 101120237) and AI4Trust (No. 101070190). We also acknowledge the support of the MUR PNRR project **iNEST**-Interconnected Nord-Est Innovation Ecosystem (ECS00000043) funded by the NextGenerationEU.

References

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015. 1

[2] Patrick PK Chan, Shute Li, Jingwen Deng, and Daniel S Yeung. Multi-proxy based deep metric learning. *Information Sciences*, 643:119120, 2023. 2, 4

[3] Dongliang Chang, Yifeng Ding, Jiyang Xie, Ayan Kumar Bhunia, Xiaoxu Li, Zhanyu Ma, Ming Wu, Jun Guo, and Yi-Zhe Song. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE TIP*, 29:4683–4695, 2020. 2

[4] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 26, 2013. 3

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6, 1

[6] Yixin Fei, Zhongkai Zhao, Siwei Yang, and Bingchen Zhao. Xcon: Learning with experts for fine-grained category discovery. In *BMVC*, 2022. 8

[7] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *ICCV*, pages 9284–9292, 2021. 2, 3, 5, 6, 8

[8] Peiyan Gu, Chuyu Zhang, Ruijie Xu, and Xuming He. Class-relation knowledge distillation for novel class discovery. In *ICCV*, pages 16474–16483, 2023. 2, 3, 6, 7

[9] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, pages 8401–8409, 2019. 2

[10] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE TPAMI*, 44(10):6767–6781, 2021. 2, 3, 6, 8

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6, 1

[12] Xuhui Jia, Kai Han, Yukun Zhu, and Bradley Green. Joint representation learning and novel category discovery on single-and multi-modal data. In *ICCV*, pages 610–619, 2021. 2

[13] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *CVPR*, pages 3238–3247, 2020. 2, 4

[14] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6

[15] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv:1610.02242*, 2016. 2

[16] Wenbin Li, Zhichen Fan, Jing Huo, and Yang Gao. Modeling inter-class and intra-class constraints in novel class discovery. In *CVPR*, pages 3449–3458, 2023. 2, 3, 6, 7

[17] Yu Liu and Tinne Tuytelaars. Residual tuning: Toward novel category discovery without labels. *IEEE TNNLS*, 2022. 2

[18] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297. Oakland, CA, USA, 1967. 5, 6, 1

[19] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, 2018. 2

[20] Zicheng Pan, Xiaohan Yu, Miaohua Zhang, and Yongsheng Gao. Ssfe-net: Self-supervised feature enhancement for ultra-fine-grained few-shot class incremental learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6275–6284, 2023. 1, 2

[21] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *CVPR*, 2023. 2

[22] Mamshad Nayeem Rizve, Navid Kardan, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Openldn: Learning to discover novel classes for open-world semi-supervised learning. In *ECCV*, pages 382–401. Springer, 2022. 3

[23] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, pages 1139–1147, 2013. 6, 1

[24] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2

[25] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, pages 7492–7501, 2022. 2, 4, 6, 8

[26] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1

[27] Jun Wang, Xiaohan Yu, and Yongsheng Gao. Feature fusion vision transformer for fine-grained visual categorization. *arXiv preprint arXiv:2107.02341*, 2021. 1, 2

[28] Muli Yang, Yuehua Zhu, Jiaping Yu, Aming Wu, and Cheng Deng. Divide and conquer: Compositional experts for generalized novel class discovery. In *CVPR*, pages 14268–14277, 2022. 2, 6

[29] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *ECCV*, pages 420–435, 2018. 2

[30] Xiaohan Yu, Yang Zhao, Yongsheng Gao, and Shengwu Xiong. Maskcov: A random mask covariance network for ultra-fine-grained visual categorization. *PR*, 119:108067, 2021. 1, 2, 3, 6

[31] Xiaohan Yu, Yang Zhao, Yongsheng Gao, Xiaohui Yuan, and Shengwu Xiong. Benchmark platform for ultra-fine-grained visual categorization beyond human performance. In *ICCV*, pages 10285–10295, 2021. 2, 6, 7

[32] Xiaohan Yu, Yang Zhao, and Yongsheng Gao. Spare: Self-supervised part erasing for ultra-fine-grained visual categorization. *PR*, 128:108691, 2022. 2, 3, 6

[33] Xiaohan Yu, Jun Wang, and Yongsheng Gao. Cle-vit: Contrastive learning encoded transformer for ultra-fine-grained visual categorization. In *IJCAI*, 2023. 2, 6, 1

[34] Xiaohan Yu, Jun Wang, Yang Zhao, and Yongsheng Gao. Mix-vit: Mixing attentive vision transformer for ultra-fine-grained visual categorization. *PR*, 135:109131, 2023. 1, 2

[35] Tong Zhang, Congpei Qiu, Wei Ke, Sabine Süsstrunk, and Mathieu Salzmann. Leverage your local and global representations: A new self-supervised learning strategy. In *CVPR*, pages 16580–16589, 2022. 2

[36] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *NeurIPS*, 34:22982–22994, 2021. 2, 3

[37] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *CVPR*, pages 10867–10875, 2021. 3, 5

[38] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *CVPR*, pages 9462–9470, 2021. 2, 3

[39] Yuehua Zhu, Muli Yang, Cheng Deng, and Wei Liu. Fewer is more: A deep graph metric learning perspective using fewer proxies. *NeurIPS*, 33:17792–17803, 2020. 2, 4