# Class-relation Knowledge Distillation for Novel Class Discovery

Peiyan Gu[1,*] Chuyu Zhang[1,2,*] Ruijie Xu[1] Xuming He[1,3]

[1]ShanghaiTech University, Shanghai, China [2]Lingang Laboratory, Shanghai, China
[3]Shanghai Engineering Research Center of Intelligent Vision and Imaging, Shanghai, China

{zhangchy2,gupy,xurj2022,hexm}@shanghaitech.edu.cn

## Abstract

*We tackle the problem of novel class discovery, which aims to learn novel classes without supervision based on labeled data from known classes. A key challenge lies in transferring the knowledge in the known-class data to the learning of novel classes. Previous methods mainly focus on building a shared representation space for knowledge transfer and often ignore modeling class relations. To address this, we introduce a class relation representation for the novel classes based on the predicted class distribution of a model trained on known classes. Empirically, we find that such class relation becomes less informative during typical discovery training. To prevent such information loss, we propose a novel knowledge distillation framework, which utilizes our class-relation representation to regularize the learning of novel classes. In addition, to enable a flexible knowledge distillation scheme for each data point in novel classes, we develop a learnable weighting function for the regularization, which adaptively promotes knowledge transfer based on the semantic similarity between the novel and known classes. To validate the effectiveness and generalization of our method, we conduct extensive experiments on multiple benchmarks, including CIFAR100, Stanford Cars, CUB, and FGVC-Aircraft datasets. Our results demonstrate that the proposed method outperforms the previous state-of-the-art methods by a significant margin on almost all benchmarks. Code is available at here.*

## 1. Introduction

The recent development of deep learning has achieved remarkable success in a broad range of visual recognition tasks [14, 13, 22]. However, most traditional methods focus on the closed-world setting, in which all the visual classes are pre-defined. As a result, it is usually difficult to deploy the learned models in realistic settings with potential novel classes. In contrast, human visual systems can efficiently acquire new concepts without supervision based on learned knowledge. Inspired by such an ability, several studies [12, 16] propose the task of Novel Class Discovery (NCD) which aims to discover novel categories from unlabeled data based on known-class data.

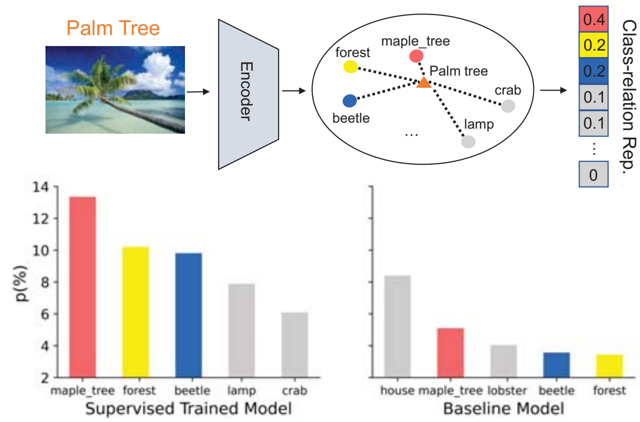


Figure 1. In the upper panel, we apply the encoder and the known class classifier to a novel sample of Palm tree, obtaining a class-relation representation. This representation encodes the relative distances between the representations of novel and known categories. The triangles indicate a novel class, while the circles show known classes. The lower panel shows the averaged class-relation representation for all Palm Tree samples and displays the 5 nearest known classes for both the *known class* Supervised Trained Model and the Baseline Model [9]. We observe that the predictions from a trained model on novel class data indicate meaningful class relation (e.g. maple tree, forest), which is lost in the Baseline Model.

A key strategy for discovering novel classes is to transfer knowledge in known classes to promote the learning of novel classes. To achieve this, most existing NCD methods [11, 31, 9] involve two training stages, including a supervised training stage followed by a discovery training stage. In the supervised training stage, they typically initialize representation by learning from known classes. In the discovery training stage, they transfer the learned knowledge to novel classes via sharing the feature representation space. While they have shown promising results, they are less effective in capturing the relationship between

[*]Both authors contributed equally.

known and novel classes, which limits the scope of shared knowledge and potentially leads to inferior representations of novel classes. Nonetheless, it is difficult to model the semantic relationship between the known and novel classes in the NCD setting as the novel classes are unknown.

To tackle this challenge, we introduce a class-relation representation for a novel class based on its similarity with the known classes. In particular, we leverage a well-known phenomenon of "dark knowledge" [15] and adopt the predicted distribution of a well-trained model to encode the inter-class relationship. To that end, we first train a model on the known classes using supervised learning and then apply the trained model to the data of novel classes. In Fig. 1 top and bottom-left, we visualize our class-relation representation and the average predictive distribution of a novel class, *palm tree*, respectively. Interestingly, we observe that the distribution often focuses on related or co-occurred classes and hence properly reflects its class relationship. For example, the *palm tree* class is closer to the *maple* and *forest* classes. However, a typical NCD baseline, which fine-tunes the pre-trained model, is unable to maintain such a similarity structure, as indicated by the example shown in Fig. 1 bottom-right. Here the *palm tree* class is more similar to the *house* class, which is less reasonable.

Motivated by the above observation, we propose a novel class-relation knowledge distillation framework for the task of novel class discovery. Our framework utilizes the class relation represented by the supervised trained model to regularize the learning of novel classes in the discovery training stage, thus preserving the meaningful class relation knowledge and promoting knowledge transfer. Moreover, to provide a flexible knowledge transfer scheme for each data sample, we develop a simple but effective learnable weight function for the regularization, which allows us to adaptively transfer knowledge based on the similarity between a novel class sample and known classes.

Specifically, we instantiate our framework with a two-head network architecture that includes an encoder and two classifier heads for the known and novel classes, respectively. We first initialize the feature representation through supervised learning on the known classes and then discover novel classes by minimizing a hybrid learning loss. Our loss consists of three terms: 1) a standard cross-entropy loss on the labeled data, which extracts semantic knowledge from the known classes; 2) an unsupervised clustering loss on the novel class data; and 3) a weighted Kullback–Leibler (KL) regularization term for distilling the class relation knowledge from the supervised trained model into the discovery of novel classes. Here the strength of each KL regularization term is controlled by a weight measuring the similarity between a novel sample and the known classes, which is derived from the predicted distribution on the known classes by the model in the discovery training stage.

To validate the effectiveness of our method, we conduct extensive experiments on four datasets, including CIFAR100, Stanford Cars, CUB, and FGVC-Aircraft. The results show that our performances surpass the previous state of the art by a large margin in most cases, demonstrating the efficacy of our novel design of learning framework. In summary, our main contributions are three-fold:

- We propose a simple and effective learning framework to facilitate knowledge transfer from the known to novel classes, which provides a new perspective to tackle novel class discovery problems.
- We propose a novel regularization strategy to capture class relation between known and novel classes via the classifier output space, and develop a simple but effective learnable weight function to adaptively transfer knowledge based on the strength of class relation.
- Our method significantly outperforms previous works on various public benchmarks, illustrating the efficacy of our design.

## 2. Related Work

**Novel class discovery:** The idea of novel class discovery was initially explored in [16, 17], which performs transfer learning across domains and tasks, and utilizes predictive pairwise similarity as the knowledge for clustering. The standard NCD problem was formalized by [12], aiming to cluster novel classes with the help of known classes. Most NCD methods attempt to transfer knowledge from known to novel classes by learning a shared representation, and can be categorized into two groups according to their clustering methods. The first group [16, 12, 31, 32, 30] typically explores pair-wise similarity for clustering. For example, KCL [16] learns a pair-wise similarity network to predict the similarity of two instances. RankStats [11] propose robust rank statistics to measure the similarity of two data in their representation space. Furthermore, DRNCD [30] presents dual rank statistics that focus on local part-level information and overall characteristics. In addition to pair-wise similarity objective, NCL [31] and Openmix [32] utilize contrastive learning and mixup strategy to promote the representation learning of novel classes. The second group [12, 31, 28] introduces self-labeling for clustering novel classes. Especially, DTC [12] utilizes deep embedding clustering to discover novel classes. UNO [9] adopts the Sinkhorn-Knopp algorithm to generate pseudo labels. Unlike above, [5] solves novel class discovery from a meta-learning perspective. Recently, several works[3, 25] have expanded the conventional novel class discovery problem to more practical scenarios where unlabeled data sets consist of known and novel classes.

Although these methods have achieved some success, few of them consider the potential relationship between the

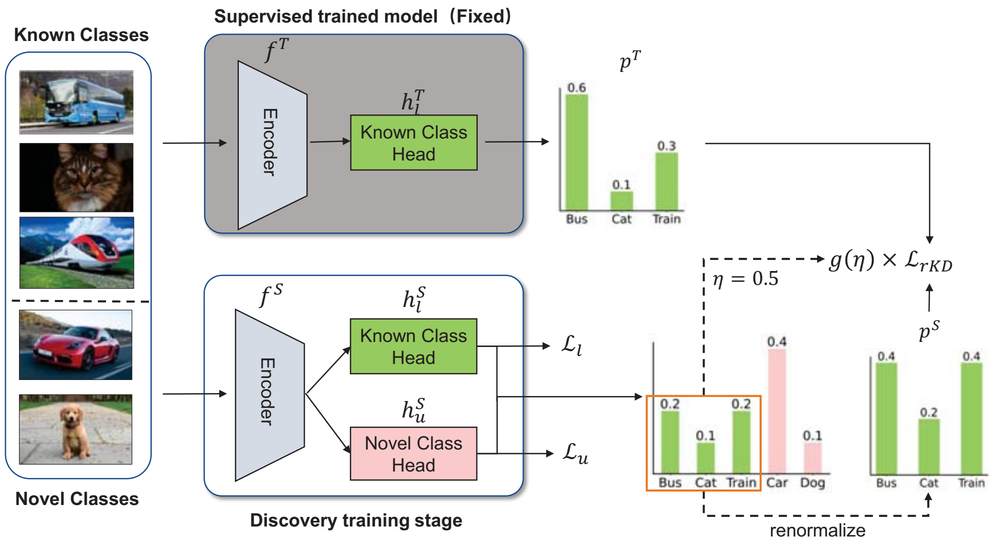


Figure 2. The overview of our class relation knowledge distillation framework. We first train a model by supervised learning on the known classes. Then we discover novel classes by jointly learning known and novel classes. To maintain meaningful relation information, we utilize the class relation represented by the supervised trained model to adaptively regularize the learning of novel classes in the discovery stage. $\eta$ represents the semantic similarity between the novel class sample and known classes, and $g(\eta)$ is the learnable weight function. And the supervised trained model is omitted in the inference.

known and novel classes in the label space during model learning. In this paper, we model the relations between known and novel classes based on model predictions, and propose a novel knowledge distillation method to transfer class-relation knowledge, thus improving the representation learning of novel classes.

**Knowledge distillation:** Knowledge distillation [15, 10, 21, 27] aims to transfer knowledge from a teacher model to a student model. Typically, it can be categorized into two groups based on whether knowledge is transferred in the model prediction space or the representation space. In the first group [15, 29], the methods often assume that the probability distribution produced by the teacher model provides more information about which classes are more similar to the predicted class than the one-hot ground truth. Therefore, they use the semantically meaningful probability distribution produced by the teacher model to supervise the learning of the student model. The second group of methods [1, 23] instead argue that the representation learned by the teacher network contains rich structural information. These methods propose to distill knowledge from the teacher to the student in the representation space by maximizing mutual information. We refer the readers to [27] for a more comprehensive survey.

In contrast to the aforementioned works, where the teacher and student networks share the same category space, our approach involves knowledge distillation between known and novel classes, which are in separate category spaces. Specifically, we distill knowledge from a model trained on known classes to a model trained on both known and novel classes, with the goal of transferring knowledge from known classes to novel classes.

# 3. Method

As shown in Fig.2, our model training is divided into two stages. In the supervised training stage, we train our model with known class data to obtain an initial feature representation, which contains meaningful semantic information, thus providing a good initialization for clustering novel classes. In the discovery training stage, we train the model with both known and novel class data, and adopt the typical cross-entropy loss and self-labeling loss to learn the known and novel classes, respectively. In addition, to better transfer knowledge, we propose a novel adaptive regularization term to distill relation knowledge from the known classes pretrained model. We will discuss the learning of our framework in detail in the following.

In this section, we first introduce our novel class relation distillation framework in Sec. 3.1. Then we describe the losses on the labeled data for the known classes and unlabeled data for novel classes in Sec. 3.2. Finally, we present our novel relation knowledge distillation loss in Sec. 3.3, which is the core design of our method.

## 3.1. Class Relation Distillation Framework

To introduce our framework, we first present the problem setting of NCD and notations. The training dataset consists of two parts: a labeled known classes set $\mathcal{D}^l = \{x_i^l, y_i^l\}_{i=0}^{|\mathcal{D}^l|}$ and an unlabeled novel classes set $\mathcal{D}^u = \{x_j^u\}_{j=0}^{|\mathcal{D}^u|}$. Here $x, y$ represent the input data and the corresponding label, respectively. We use $Y^l = \{1, 2, ..., C^l\}$ and $Y^u = \{C^l + 1, C^l + 2, ..., C^l + C^u\}$ to represent the category space of known and novel classes, respectively.

We adopt a common model architecture for NCD, consisting of an encoder, denoted by $f$, along with two cosine classifier heads: $h_l$ for known classes and $h_u$ for novel classes. The encoder can be a standard convolutional network (CNN) or Vision Transformer (ViT) [8]. Given an input image from a known or novel class, we first project it into an embedding space through the shared encoder. Then we normalize the embedding and feed it to the known and novel class head. Note that no matter whether an input is known or novel, it will go through two heads to generate two outputs. Finally, we concatenate the two outputs as the final prediction. The forward process can be written as:

$$p(y|x) = \text{Softmax}((h_l^S(f^S(x)) \oplus h_u^S(f^S(x)))/\tau) \quad (1)$$

where superscript $S$ denotes the model in discovery training stage, $p(y|x) \in \mathbb{R}^{C^l+C^u}$ is the model predictive distribution, and $\tau$ is the temperature of the softmax function.

To discover novel classes, we begin by initialize our representation ability using supervised learning with known classes, then discover novel class by training on known and novel class data jointly. In the discovery training stage, our objective function consists of three terms: 1) a supervised loss for known class data, 2) an unsupervised loss for novel class data, and 3) a class-relation Knowledge Distillation loss for novel class data. The overall loss can be written as:

$$\mathcal{L} = \mathcal{L}_l + \alpha\mathcal{L}_u + \beta\mathcal{L}_{rKD} \quad (2)$$

where $\mathcal{L}_l$ is the standard supervised loss on known classes data, $\mathcal{L}_u$ is the unsupervised clustering loss for novel classes data, and $\mathcal{L}_{rKD}$ is our relation Knowledge Distillation loss. Here $\alpha, \beta$ are the weighting factors.

### 3.2. Loss for known and novel classes

We now present the first two loss terms for the known and novel class data in the discovery training stage, respectively. For the supervised loss on the known classes, we adopt the standard cross-entropy loss. For the unsupervised clustering loss on the novel classes data, we adopt the widely used self-labeling loss [2, 9], which assigns pseudo labels for novel classes data by solving an optimal transport (OT) problem, then utilizes the generated pseudo label to self-train the model.

Specifically, such a self-labeling process assumes that the data of novel classes are equally partitioned into clusters and utilizes Sinkhorn-knopp algorithm to find an approximate assignment. We denote $\mathbf{y}^q = q(y^u|x^u)$ as pseudo label, $\mathbf{y}^p = p(y^u|x^u)$ as model's prediction, and $\mathbf{y}^p, \mathbf{y}^q \in \mathbb{R}^{C^u \times 1}$. Let $\mathbf{Q} = [\mathbf{y}_1^q, \mathbf{y}_2^q, , , \mathbf{y}_B^q]\frac{1}{B}$, $\mathbf{P} = [\mathbf{y}_1^p, \mathbf{y}_2^p, , , \mathbf{y}_B^p]\frac{1}{B}$ be the joint distribution of $B$ sampled data. We estimate $\mathbf{Q}$ by solving an OT problem:

$$\langle \mathbf{Q}, -\log \mathbf{P} \rangle_F$$

$$\text{s.t. } \mathbf{Q} \in \{\mathbf{Q} \in \mathbb{R}_+^{C^u \times B} | \mathbf{Q}\mathbf{1}_B = \frac{1}{C^u}\mathbf{1}_{C^u}, \mathbf{Q}^\top \mathbf{1}_{C^u} = \frac{1}{B}\mathbf{1}_B\}$$

where $\langle , \rangle_F$ is the Frobenius inner product. We refer readers to [6, 2] for the details of optimization. The optimal $\mathbf{Q}$ is the pseudo label of unlabeled data and we denote the optimal pseudo label as $q^*(y^u|x^u)$. The self-labeling loss is:

$$\mathcal{L}_u = \frac{1}{B}\sum_{i=1}^{B} -q^*(y_i^u|x_i^u) \log p(y_i^u|x_i^u) \quad (3)$$

To transfer knowledge between known and novel classes, previous methods [12, 30, 3] couple the learning of known and novel classes by sharing the encoder $f^S$. The models are typically learned by optimizing the supervised cross-entropy loss on labeled data and the self-labeling loss on unlabeled data. This parameter sharing and the jointly-optimized model allow them to learn representations helpful for novel class clustering. However, such an implicit knowledge transfer method is incapable of fully utilizing the knowledge contained in known classes for the clustering of novel classes. Below, we introduce a novel class relation knowledge distillation term to constrain the model learning in the discovery training phase, resulting in a better representation and learning of the novel classes.

### 3.3. Class-relation Knowledge Distillation

For more effective knowledge transfer, we introduce a class-relation representation based on the output distribution of a model on the known classes classifier. Such a distribution encodes the similarity structure between a novel class data and the known classes. However, current methods transfer knowledge by sharing an encoder, which is less effective in capturing class relations for novel classes. In particular, as shown in Fig. 1, we find that there is a meaningful class relation contained in the supervised pretrained model, but the class relation is less meaningful for the baseline model. We speculate that such a class relation is important for learning a good representation of novel classes, while the conventional discovery training stage inadvertently changes the representation space and weakens the relations between known and novel classes.

Therefore, we propose a novel relation Knowledge Distillation (rKD) loss to distill the knowledge contained in the supervised trained model to enhance the learning of novel classes. Our rKD loss regularizes the model learning process in the discovery stage, preventing the model from losing meaningful class relations. Moreover, since the class relation may vary for different novel class samples, we consider an adaptive regularization scheme based on the strength of the class relation. For instance, for novel class samples that are more similar to known classes, we impose a larger regularization weight. To achieve this, we propose a simple but effective learnable weight function to control the regularization effect for different novel samples. In the following section, we formally introduce our novel rKD loss and weight function.

**Knowledge Distillation:** In the discovery training stage, for novel classes, we encourage the model to learn discriminative representation while maintaining the relations to the known classes. To this end, we keep the supervised trained model and feed each novel-class sample into the model to compute an initial relation representation. Similarly, we use the model $f^S$ to obtain the current class relation. This process can be written as follows:

$$p^T = \text{softmax}((h_l^T(f^T(x^u)))/t) \in \mathbb{R}^{C^l} \tag{4}$$

$$p^S = \text{softmax}((h_l^S(f^S(x^u)))/t) \in \mathbb{R}^{C^l} \tag{5}$$

where $p^T, p^S$ denotes the relation representation of the supervised and discovery-stage model respectively. $t$ is the temperature used to soften relation representation. A higher value in $p^T$ means the novel class sample is semantically closer to the corresponding known class. To regularize the model learning in the discovery training stage, we propose a novel knowledge distillation (KD) loss term based on the KL divergence between $p^S$ and $p^T$:

$$\mathcal{L}_{rKD}^u = \frac{1}{B}\sum_{i=1}^{B} \text{KL}(p_i^T || p_i^S) \tag{6}$$

$p^T$ is fixed in the discovery training stage. The loss $\mathcal{L}_{rKD}^u$ regularizes the learned representations to maintain relative relation between novel class samples and known classes represented by the supervised trained model.

**Learnable weight function:** Typically, novel-class samples have varying semantic similarities and for the semantic dissimilar samples, the relation represented by $p^T$ is relatively noisy. This requires an adaptive regularization for the novel-class data. To tackle this, we propose a simple but effective learnable weight function to control the regularization strength for different novel class samples. For a novel class sample, we first utilize the sum of known classes' probability in Eq.1 to represent the relation strength to the known classes. Formally, the relation strength for the ith sample in a batch is:

$$\eta_i = \sum_{k=1}^{C^l} p_{i,k}(y_i^u | x_i^u) \tag{7}$$

where $p_{i,k}(y_i^u|x_i^u)$ denotes the sample $x_i^u$ probability on class $k$. Higher $\eta$ indicates a stronger semantic relation with the known classes. Then, we develop a learnable weight function $g$ as a positive correlation function about $\eta$. In this work, we adopt a simple design that computes a normalized relation strength over the batch:

$$g(\eta_i) = \text{Norm}(\eta_i) = B\frac{\eta_i}{\sum_{j=1}^{B}\eta_j} \tag{8}$$

In our experiments, the batch size is large enough to ensure that the statistics of the mean are stable. With our learnable weight function, the adaptive relation knowledge distillation loss for novel classes can be written as:

$$\mathcal{L}_{rKD} = \frac{1}{B}\sum_{i=1}^{B} g(\eta_i)\text{KL}(p_i^T || p_i^S) \tag{9}$$

With our novel adaptive class relation knowledge distillation regularization term, our model can cluster novel classes and maintain the semantic meaningful representation structure simultaneously.

The utilization of this proposed design confers three advantages. Firstly, the function $g(\eta)$ exhibits a positive correlation with $\eta$, thereby directing the model to prioritize the KL loss of samples that share a higher similarity with the known classes. Secondly, given that the mean value of $\eta$ may vary across datasets, the normalization procedure ensures that the weight values $g(\eta)$ remain uniform across datasets, leading to a more consistent application of the hyperparameter of $\mathcal{L}_{rKD}$. Finally, the learnability of our weight function allows the weight function to adapt dynamically based on the relation's learning dynamics. In particular, for samples with a higher KL divergence, we optimize relations by simultaneously decreasing the KL divergence term and weight function to downweight the learning on challenging examples. Conversely, for samples with a lower KL divergence, we relatively increase the weight function, empowering the model to learn shared semantic information between the novel and known classes more effectively. We ablate the design of the weight function in the experiments.

Table 1. The details of dataset split.

| Dataset | Known | | Novel | |
|---|---|---|---|---|
| | Images | Classes | Images | Classes |
| CIFAR100-20 | 40.0k | 80 | 10.0k | 20 |
| CIFAR100-50 | 25.0k | 50 | 25.0k | 50 |
| Stanford Cars | ≈4.0k | 98 | ≈4.1k | 98 |
| CUB | ≈3.0k | 100 | ≈3.0k | 100 |
| FGVC-Aircraft | ≈3.3k | 50 | ≈3.3k | 50 |

# 4. Experiments

## 4.1. Experimental Setup

**Datasets:** To evaluate the effectiveness of our method, we first conduct tests on the typical CIFAR100 dataset [19]. Specifically, we divide CIFAR100 into two categories: 80/20 known/novel classes and 50/50 known/novel

Table 2. Comparison with the SOTA methods on the unlabeled training set of the CIFAR100, Stanford Cars, CUB, and Aircraft datasets.

| Method | CIFAR100-50 | CIFAR100-80 | Stanford Cars | CUB | Aircraft |
|---|---|---|---|---|---|
| Kmeans | 28.3±0.7 | 56.3±1.7 | 13.1±1.0 | 42.2±0.5 | 18.5±0.3 |
| DTC[12] | 35.9±1.0 | 67.3±1.2 | - | - | - |
| RankStats+[11] | 44.1±3.7 | 75.2±4.2 | 36.5±0.6 | 55.3±0.8 | 38.4±0.6 |
| NCL[31] | 52.7±1.2 | 86.6±0.4 | 43.5±1.2 | 48.1±0.9 | 43.0±0.5 |
| ComEx[28] | 53.4±0.7 | 85.7±1.3 | - | - | - |
| UNO[9] | 60.4±1.4 | 90.4±0.2 | 49.8±1.4 | 59.2±0.4 | 52.1±0.7 |
| GCD[25] | - | - | 42.6±0.4 | 56.4±0.3 | 49.5±1.0 |
| Ours | **65.3**±0.6 | **91.2**±0.1 | **53.5**±0.8 | **65.7**±0.6 | **55.8**±0.9 |

classes, with the latter being more challenging. As the results on CIFAR10 [19] and ImageNet[7] datasets are nearly saturated [9], we turn to evaluate our method on three fine-grained datasets - Stanford Cars [18], CUB [26], and FGVC-Aircraft [20] - which are more demanding for novel class discovery. We divide these datasets into two halves, with one comprising known classes and the other consisting of novel classes. The details of the dataset splits are presented in Tab.1.

**Metric:** Following [9], we evaluate our method in the transductive learning setting and inductive learning setting. In the transductive learning setting, we employ ClusterAcc to evaluate the train novel datasets. The formula for ClusterAcc is as follows:

$$\text{ClusterAcc} = \max_{perm \in P} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{y_i = perm(\hat{y}_i)\} \quad (10)$$

where $y_i$ and $\hat{y}_i$ denote the ground-truth and predicted labels, respectively, while $P$ represents the set of all permutations. We use the Hungarian algorithm to find the best permutation. In the inductive learning setting, we use the task-agnostic evaluation protocol [9] to evaluate model performance. The accuracy of the labeled data is calculated by using the standard accuracy metric. For unlabeled data, we use ClusterAcc to measure the performance of class discovery. This protocol is applied to a test set that includes both known and novel class data. As this evaluation does not require any prior knowledge of whether the data is novel or known, it is more suitable for real-world applications than the first type of evaluation.

**Implementation Details:** For the CIFAR100 dataset, following previous work [11, 9], we use ResNet18 as our backbone network. First, we pretrain the network on the known classes for 100 epochs, and then we jointly train our network on the known and novel classes for 500 epochs, which is similar to UNO [9]. The optimizer is SGD, and the learning rate first grows linearly and then cosine decays. We implement NCL [31] on the CIFAR100-50 setting, and other results are mostly cited from their papers. For Stanford Cars, CUB, Aircraft datasets, we utilize DINO [4] pretrained ViT as our backbone, and we only finetune the last block of ViT by Adamw optimizer. For a fair comparison, we adopt the same training setting to implement RankStats [11], NCL [31] and UNO [9] on those datasets based on their released code, and the training is all converged. Specially, we first pretrain the network on the known classes for 50 epoch, and then jointly train the network on the known and novel classes for 100 epochs. Regarding hyperparameters, we adhere to the settings outlined in [9] and [15], assigning a value of 0.1 to $\tau$, 1 to $\alpha$ and 0.4 to $t$ for all datasets. Additionally, we set $\beta$ to 0.1 for most datasets, except CIFAR100-80, which we set to 0.02. Furthermore, we analyze its sensitivity in the experiments. More comprehensive details are in Appendix.

### 4.2. Results

**Comparison with SOTA:** We compare our method with currently state-of-the-art, including UNO [9], ComEx [28], NCL [31] and RankStats+ [11] in Tab.2. Following them, we first present the results on the typical CIFAR100 dataset. Our result on the CIFAR100-50 dataset significantly outperforms SOTA. In particular, we outperform UNO by 4.9%. On three challenging fine-grained datasets, our experimental results demonstrate that our proposed approach outperforms UNO [9] by 3.7%, 6.5%, and 3.7% on Stanford Cars, CUB, and Aircraft datasets, respectively.

In addition, as shown in Tab.3, we report the results on the test dataset with task-agnostic evaluation metric. Overall, our method outperforms previous methods by a significant margin on both known and novel classes. Specifically, on the typical CIFAR100-50 dataset, we achieve 3.6% and 1.8% improvement on known and novel classes, respectively. On three challenging fine-grained datasets, our method outperforms the previous state-of-the-art (UNO) by 3-5% on novel classes, while also obtaining 1-2% gains on known classes. We hypothesize that our method enhances the learning of novel classes by leveraging class relationships, resulting in a direct improvement in the clustering effectiveness of novel classes, as well as a reduction in noise during the learning process. This, in turn, enhances representation learning and indirectly improves the performance

Table 3. Comparison with state-of-the-art methods on CIFAR100, Stanford Cars, CUB, and Aircraft datasets under the inductive setting, using task-agnostic evaluation protocol. GCD [25] is not applicable to the test set.

| Method | CIFAR100-50 | | | Stanford Cars | | | CUB | | | Aircraft | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Known | Novel | All | Known | Novel | All | Known | Novel | All | Known | Novel | All |
| RankStats+ [12] | 69.7 | 40.9 | 55.3 | 81.8 | 31.7 | 56.3 | 80.7 | 51.8 | 66.1 | 66.4 | 36.5 | 51.5 |
| NCL [31] | 72.4 | 25.7 | 49.0 | 83.5 | 24.4 | 53.4 | 79.8 | 13.1 | 46.3 | 62.8 | 26.5 | 44.6 |
| UNO [9] | 75.0 | 57.6 | 66.3 | 81.7 | 46.7 | 63.9 | 78.7 | 62.1 | 70.3 | 71.2 | 52.4 | 61.8 |
| Ours | **78.6** | **59.4** | **69.0** | **83.9** | **51.3** | **67.3** | **81.1** | **67.5** | **74.2** | **72.2** | **55.2** | **63.7** |

of known classes. In conclusion, the strong performance of our method on those challenging datasets demonstrates the effectiveness of our approach.

**The number of clusters is unknown:** The above experiments assume that the number of novel classes is known. However, this is unrealistic in practice. Therefore, in order to further validate the effectiveness of our method in practical scenarios, we conduct experiments in the case of an unknown number of classes. We assume that the deviation between the estimated classes and the true classes is between -20% and 20%. For the case where there are 100 novel classes, -20% means that the estimated classes are 20 less than the true classes, and +20% means that the estimated classes are 20 more than the true classes. We report the results on the train novel dataset. As shown in Fig.3, in most cases, our method performs better than previous methods. Moreover, the more accurate the class estimation, the more obvious the advantage of our method. We speculate that when the class estimation error is relatively large, multiple novel classes may merge or split, making the relationship between the novel and known classes noisy and not conducive to learning relations.

**Visualization:** We conducted a qualitative analysis of the learned feature space using t-SNE [24]. As depicted in Fig.4, the supervised pre-trained model is noisy and some classes produced by UNO [9] are entangled, making it difficult for a linear classifier to distinguish the samples. In contrast, our proposed method generates more compact feature representations that tightly group samples of the same class.

We also present the class relationships produced by our model on CIFAR100-50. Fig.5 illustrates the averaged predictions for instances of the novel "ray" class on the known class head. It shows that the baseline method fails to capture the "dolphin $>$ flatfish" relation order of "ray", and while our proposed method preserves this order. Additionally, our approach exhibits superior performance in filtering out background noise, such as "cloud" and "mountain", and discovering innovative relationships between categories, such as "aquarium fish" and "lizard". More visualizations are available in the appendix.

Overall, our findings demonstrate the effectiveness of our approach in learning feature representations and class relationships.

Table 4. Ablation study. $\mathcal{L}^u_{rKD}$ stands for adding KD loss on the unlabeled data, and $g(\eta)$ is the learnable weight function on $\mathcal{L}^u_{rKD}$. All results are evaluated on the unlabeled training set.

| $\mathcal{L}^u_{rKD}$ | $g(\eta)$ | Stanford Cars | CUB | Aircraft |
|---|---|---|---|---|
| ✗ | ✗ | 49.1 | 59.2 | 52.1 |
| ✓ | ✗ | 51.9 | 63.7 | 53.4 |
| ✓ | ✓ | **53.5** | **65.7** | **55.8** |

Table 5. Ablation the design of our learnable weight function $g(\eta)$. SG and Norm denote Stop Gradient and Normalization, respectively. All results are evaluated on the unlabeled training set.

| $g(\eta)$ | Stanford Cars | CUB | Aircraft |
|---|---|---|---|
| 1 | 51.9 | 63.7 | 53.4 |
| $\eta$ | 32.1 | 58.5 | **58.5** |
| SG($\eta$) | 51.2 | 64.3 | 54.7 |
| SG(Norm($\eta$)) | 51.9 | 64.6 | 54.2 |
| Norm($\eta$) | **53.5** | **65.7** | 55.8 |

## 4.3. Ablation study

**Component Analysis:** As shown in Tab. 4, we conduct an ablation study to evaluate the effectiveness of our novel class relation knowledge distillation regularization ($\mathcal{L}^u_{rKD}$) and learnable weight function ($g(\eta)$). With $\mathcal{L}^u_{rKD}$, we see improvements of 2.8%, 4.5%, and 1.3% on the Stanford Cars, CUB, and Aircraft datasets, respectively, demonstrating that the class relation helps guide the learning of novel classes. Furthermore, the incorporation of the learnable weight function leads to even further improvement in all datasets, confirming the effectiveness of both modules.

**Learnable weight function:** Our goal is to apply constraints with different intensities to samples that have varying semantic relationships. Specifically, we apply stronger constraints to samples with higher semantic relationships.

In Table 5, we investigate several simple designs for the weight function and analyze them from the perspectives of stop gradient (SG) and normalization (Norm). For instance, SG($\eta$) involves using $g(\eta) = \eta$ without backpropagating gradients. We compare Norm($\eta$) with SG(Norm($\eta$)), yielding significant improvement on all three datasets, indicating

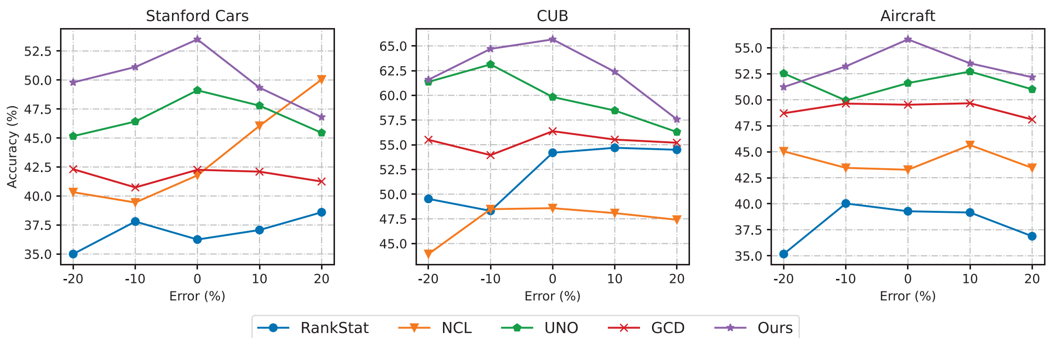


Figure 3. The x-axis in these plots represents the error rate of the estimated number of novel clusters. For the CUB dataset, which has 100 novel classes, -20% and 20% denote underestimated and overestimated 20 classes, respectively.

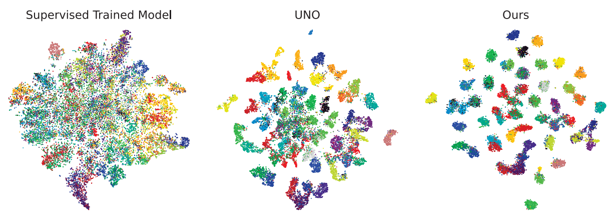


Figure 4. t-SNE visualization of unlabeled training set on CIFAR100-50.

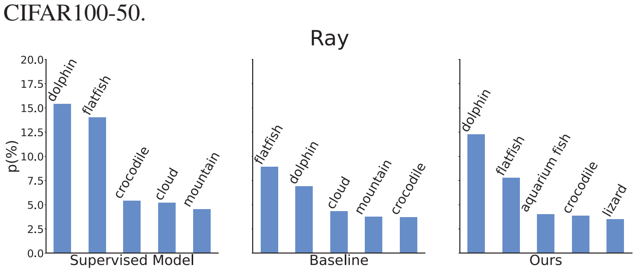


Figure 5. Visualization of quantified relative relationships. The bar labels represent the known classes in the CIFAR100-50 setting. Each plot shows average predictions for instances of the novel "ray" class on the known class head.

the superiority of the learnable characteristics. Moreover, Norm($\eta$) is more stable than $\eta$, and outperforms $\eta$ on Stanford Cars and CUB datasets, validating the effectiveness of normalization. Although there is theoretically a degenerate solution in our weight function, which assigns the maximum weight to the sample with the smallest KL loss term. In practice, due to the randomness of batch samples, the model is difficult to optimize towards this degenerate solution. More analysis is in the Appendix.

**Hyperparameter $\beta$:** Our proposed approach is simple and effective, utilizing a single hyperparameter $\beta$ to regulate the impact of the class relation regularization term. As shown in Tab.6, our experiments on the unlabeled training dataset (Train-Novel) demonstrate that various values of $\beta$ result in significant improvements over the baseline model with $\beta = 0$. Furthermore, we observe that as the value of $\beta$ increases, the relative gains become increasingly conspicuous. However, it is important to note that excessively high

Table 6. Analysis of hyperparameter $\beta$. "Train-novel" refers to the evaluation of an unlabeled training dataset, whereas the other results are evaluations of the test set.

| $\beta$ | | CUB | | |
|---|---|---|---|---|
| | Train-Novel | Known | Novel | All |
| 0 | 59.2 | 78.7 | 62.1 | 70.3 |
| 0.01 | 62.1 | 80.5 | 63.1 | 71.8 |
| 0.02 | 62.0 | 80.4 | 64.0 | 72.2 |
| 0.05 | 63.6 | 81.0 | 66.5 | 73.7 |
| 0.1 | 65.7 | **81.0** | 67.5 | **74.2** |
| 0.2 | 65.1 | 80.3 | 65.8 | 73.0 |
| 0.5 | 67.3 | 78.3 | 67.4 | 72.8 |
| 1 | **67.9** | 77.2 | **68.4** | 72.8 |

values of $\beta$ may produce a model that overemphasizes novel class learning at the expense of weaker performance in the known classes. This, in turn, can lead to lower overall performance during testing. Based on our results, we suggest using a default value of $\beta = 0.1$, which achieves a better balance between known and novel class learning.

## 5. Conclusion

In this paper, we propose a novel class-relation knowledge distillation learning framework, which provides a new perspective to transferring knowledge from known to novel classes in the NCD problem. Instead of transferring knowledge only by sharing representation space, we utilize class relations to transfer knowledge. Specifically, we observe that the prediction distribution of novel classes on a model trained on known classes effectively captures the relationship between the novel and known classes. However, this relationship is disrupted during the discovery training stage. Therefore, to maintain this meaningful inter-class relationship, we propose a simple and effective regularization term that constrains the model in the discovery training stage. Additionally, we propose a learnable weight function that dynamically assigns more weight to semantically similar samples, enabling the model to learn the shared semantic information. Our method achieves significant improvements on several general datasets and fine-grained datasets, validating the effectiveness of our approach. Furthermore, we hope our findings will shed more light on future work to explore the relationship between known and novel classes and enhance the model's transferbility for knowledge transfer.


## Acknowledgement

This work was supported by Shanghai Science and Technology Program 21010502700, Shanghai Frontiers Science Center of Human-centered Artificial Intelligence and MoE Key Lab of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University).


## References


[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9163–9171, 2019.

[2] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[3] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. *International Conference on Learning Representations (ICLR)*, 2022.

[4] 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.

[5] Haoang Chi, Feng Liu, Wenjing Yang, Long Lan, Tongliang Liu, Bo Han, Gang Niu, Mingyuan Zhou, and Masashi Sugiyama. Meta discovery: Learning to discover novel classes given very limited data. In *International Conference on Learning Representations*, 2022.

[6] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[7] 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.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] 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.

[10] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[11] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[12] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8401–8409, 2019.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[16] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *International Conference on Learning Representations*, 2018.

[17] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *International Conference on Learning Representations*, 2018.

[18] 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.

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[20] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[21] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[23] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *International Conference on Learning Representations*, 2020.

[24] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[25] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. *arXiv preprint arXiv:2201.02609*, 2022.

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

[27] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[28] Muli Yang, Yuehua Zhu, Jiaping Yu, Aming Wu, and Cheng Deng. Divide and conquer: Compositional experts for generalized novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14268–14277, 2022.

[29] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022.

[30] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *Advances in Neural Information Processing Systems*, 34, 2021.

[31] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10875, 2021.

[32] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9462–9470, 2021.