OpenMix: Reviving Known Knowledge for Discovering Novel Visual Categories in an Open World

Anonymous CVPR 2021 submission

Paper ID 866

Abstract

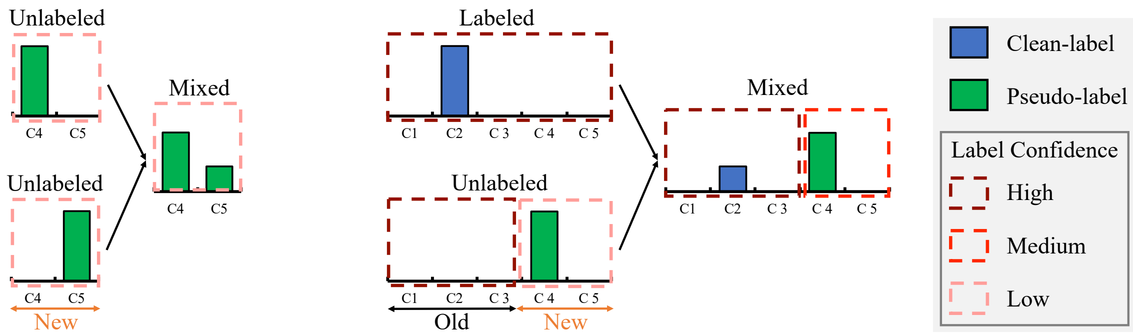
In this paper, we tackle the problem of discovering new classes in unlabeled visual data given labeled data from disjoint classes. Existing methods typically first pre-train a model with labeled data, and then identify new classes in unlabeled data via unsupervised clustering. However, the labeled data that provide essential knowledge are often underexplored in the second step. The challenge is that the labeled and unlabeled examples are from non-overlapping classes, which makes it difficult to build the learning relationship between them. In this work, we introduce OpenMix to mix the unlabeled examples from an open set and the labeled examples from known classes, where their non-overlapping labels and pseudo-labels are simultaneously mixed into a joint label distribution. OpenMix dynamically compounds examples in two ways. First, we produce mixed training images by incorporating labeled examples with unlabeled examples. With the benefit of unique prior knowledge in novel class discovery, the generated pseudo-labels will be more credible than the original unlabeled predictions. As a result, OpenMix helps preventing the model from overfitting on unlabeled samples that may be assigned with wrong pseudo-labels. Second, the first way encourages the unlabeled examples with high class-probabilities to have considerable accuracy. We introduce these examples as reliable anchors and further integrate them with unlabeled samples. This enables us to generate more combinations in unlabeled examples and exploit finer object relations among the new classes. Experiments on three classification datasets demonstrate the effectiveness of the proposed OpenMix, which is superior to state-of-the-art methods in novel class discovery.

1. Introduction

Recent advances in deep learning have witnessed great developments in visual recognition, especially image classification [4]. It is reported that modern image classification models [17, 9] can identify thousands of classes with high accuracy, but require a large number of labeled training samples. Although semi-supervised learning (SSL) [2] and noisy label learning (NLL) [6] can mitigate the need for annotations and maintain high performance, they still require some clean (SSL) or noisy (NLL) annotations for every class of interest. Furthermore, the generalization ability of learned classifiers is far from the human ability. In fact, a human can easily identify samples of new classes that may appear in real applications. However, a learned classifier can only recognize samples of the known classes, but is likely to fail handling the ones of unseen (new) classes. That is, it is significantly difficult and still underexplored to identify new classes that are undefined previously and do not have any annotated samples.

In this work, we attempt to address the recent proposed problem, called novel class discovery [8], where we are given labeled data of known (old) classes and unlabeled data of novel (new) classes. It is an open set problem where classes of unlabeled data are undefined previously and annotated samples of these novel classes are not available. The goal of novel class discovery is to identify new classes in unlabeled data with the support of knowledge of old classes. To achieve this objective, existing methods [7, 8, 10, 11] commonly follow a two-step learning strategy: 1) pre-train the model with labeled data to obtain basic discriminative ability; 2) recognize new classes in unlabeled data via unsupervised learning upon the trained model. However, the labeled data are only used to learn off-the-shelf features in the first step, but are largely ignored in the second step. In this way, the model can only benefit from the off-the-shelf knowledge of the labeled data, but fails to leverage the underlying relationship between the labeled and unlabeled data. In this work, we argue that the labeled data provide essential knowledge about underlying object structures and common visual patterns. However, the use of labeled data is much harder than in semi-supervised learning [2, 15], due to the fact that the labeled and unlabeled samples are from disjoint classes.

To this end, the question is *how to effectively exploit the*



(a) MixUp among unlabeled samples (b) OpenMix among labeled and unlabeled samples

Figure 1. Examples of (a) directly using MixUp among unlabeled samples and (b) the proposed OpenMix. Due to the uncertainty of pseudo-labels of unlabeled samples, their mixed labels may still have *low* confidence. In OpenMix, the prior knowledge (area of *high* confidence) leads the mixed label to have *high* (exactly true) confidence in old classes and *medium* (reliable) confidence in new classes.

labeled data to promote the discovery of new classes? In this work, we try to answer this question and propose a simple but effective method, called OpenMix, for the open set problem considered in this paper. OpenMix is largely motivated by MixUp [29], which is widely used in supervised learning [29, 27] and semi-supervised learning [2, 1]. However, one premise of using MixUp is that there should be labeled samples for every class of interest, which is not appropriate for our task. This is because we only have pseudo-labels for unlabeled samples of new classes, and the accuracy of these pseudo-labels can not be guaranteed. If we directly apply MixUp on unlabeled samples along with their uncertain pseudo-labels, the generated pseudo-labels will still be unreliable (Fig .1 (a)). Training with these unreliable pseudo-labels may further damage the model performance. Therefore, it is non-trivial to adopt MixUp for novel class discovery.

Instead of readily using MixUp on unlabeled samples, during the unsupervised clustering, OpenMix generates training samples by incorporating both labeled and unlabeled samples. OpenMix compounds samples in two ways. First, OpenMix mixes the labeled samples with unlabeled samples. Meanwhile, since the labeled and unlabeled samples belong to different label spaces, we first extend their labels/pseudo-labels to joint label distributions, and then mix them. OpenMix leverages two priors in novel class discovery: 1) labels of labeled samples of old classes are exactly clean, and 2) labeled and unlabeled samples belong to completely different classes. These two properties encourage the pseudo-labels of mixed samples to have 1) exactly true confidence in old classes and 2) higher confidence in new classes (Fig .1 (b)). That is, in the old class set, the pseudo-label of a mixed sample is correct, because the label of the labeled counterpart is correct and the unlabeled counterpart does not belong to any old classes. On the other hand, in the new class set, the uncertainty of a pseudo-label will be partially eliminated by mixing with the labeled sample. This is because the labeled counterpart does not belong to any new classes and its label distribution in the new class set is exactly true. With the above properties, the pseudo-labels of mixed samples will be more reliable than those of their unlabeled counterparts. As a result, OpenMix can help preventing the model from overfitting on unlabeled samples that may be assigned wrong pseudo-labels. Second, we observe that the first way of OpenMix encourages the model to keep high classification accuracy for unlabeled samples having high class-probabilities. Therefore, we select these samples as reliable anchors of new classes and mix them with unlabeled samples for further improvement.

In summary, the contributions of this paper are:

- This work proposes the OpenMix, which is tailor-made for effectively leveraging known knowledge in novel class discovery. OpenMix can prevent the model from fitting on wrong pseudo-labels, thereby consistently improving the model performance.
- OpenMix enables us to explore reliable anchors from unlabeled samples, which can be used to generate diverse smooth samples of new classes towards a more discriminative model.
- This paper presents a simple baseline for novel class discovery, which can achieve competitive results.

Experiments conducted on three datasets show that our approach outperforms the state-of-the-art methods by a large margin in novel class discovery.

2. Related Work

This work is related to novel class discovery, unsupervised clustering, transfer learning, semi-supervised learning and MixUp [29]. We briefly review the most representative works and discuss the relationship with them.

Novel Class Discovery is a recent task aiming at recognizing novel classes in unlabeled data. Different from the traditional unsupervised learning, this task also provides labeled data of other classes. Existing methods usually use the labeled data for model initialization and perform unsupervised clustering on unlabeled data. In [10] and [11], a Constrained Clustering Network (CCN) is proposed. CCN first trains a binary-classification model on labeled data to estimate pair-wise similarity of images. Then, a clustering model is trained on unlabeled data by using the prediction of the binary-classification model as supervision. The difference between [10] and [11] is the loss function used in CCN. Han *et al.* [8] first pre-train the model on labeled data by cross-entropy loss and then implement clustering on unlabeled data by DEC [25]. Latter, Han *et al.* [7] propose employing rank statistics to estimate the pairwise similarity of images. The pairwise pseudo-labels are used to achieve unsupervised clustering on the unlabelled data. Except [7], none of the above methods use the labeled data during the stage of unsupervised clustering. In [7], the labeled data are mainly used to keep the model accuracy on old classes. By contrast, our goal is improving the accuracy on new classes with the labeled data.

Unsupervised Clustering focuses on automatically dividing unlabeled data. Many classic methods [14, 28] and deep learning methods [25, 5, 26] have been proposed. Unlike novel class discovery, there is no prior knowledge provided (*e.g.*, labeled data) for unsupervised clustering. In such a context, there may be multiple criteria for most datasets, such as color, shape, and other attributes, so that the clustering results may not fit the expectation. In contrast, the labeled data in novel class discovery provide useful knowledge and can guide us to learn clustering models that match the clustering criteria of labeled data.

Transfer Learning [16, 23, 19] aims to transfer the knowledge of a labeled dataset to another dataset. Generally, the classes of the new (target) dataset are different from the previous (source) one. In transfer learning, both the source and target data are labeled. Instead, the target data are unlabeled in novel class discovery, leading this task to be more difficult.

Semi-Supervised Learning [20, 15, 2, 1] is designed to training a model on a partially labeled dataset. Novel class discovery is similar to this task in that both tasks are provided with labeled and unlabeled samples. The difference is that the labeled and unlabeled samples share the same class set in semi-supervised learning. However, the classes of labeled and unlabeled samples are completely different in novel class discovery.

MixUp [29] has been utilized successfully in supervised learning [27, 29] and semi-supervised learning [1, 2]. Unlike existing MixUp-based methods, we apply MixUp to effectively leverage labeled data of known classes for novel class discovery. In addition, existing MixUp-based methods assume that there are some clean labels for every class of interest, which is an important precondition. However, in novel class discovery, the labels of new classes are not available, causing MixUp not to be directly applicable without careful design.

3. Our Method

In novel class discovery, we are provided with labeled data $D^l = \{X^l, Y^l\}$ and unlabeled data $D^u = \{X^u\}$. The number of samples is N^l in D^l and N^u in D^u, respectively. Each labeled image x_i^l has a label y_i^l, where $y_i^l \in \{1, 2, ..., C^l\}$ and C^l is the number of classes of D^l. Following [7], we assume the number of classes of D^u is prior knowledge, which is defined as C^u. The classes of D^l and D^u are disjoint. We define the classes as old classes and new classes for D^l and D^u, respectively. The goal of novel class discovery is to leverage the knowledge of D^l to identify the classes in D^u.

In this paper, we try to achieve this goal by learning a model constructed by a convolutional neural network (CNN) and two classifiers. These two classifiers, called old classifier and new classifier, are used to recognize samples from old classes and new classes, respectively. The framework of our method is illustrated in Fig. 2. Next, we will present the baseline model for novel class discovery.

3.1. Baseline

In this work, we follow the two-stage learning strategy to design the baseline. In the first stage, we utilize the labeled data to train the CNN and the old classifier, which can provide basic discriminative representations for images and accurately classify samples of old classes. In the second stage, we learn an unsupervised clustering model on the unlabeled data by pseudo-pair learning and pseudo-label learning, enabling us to identify samples of new classes.

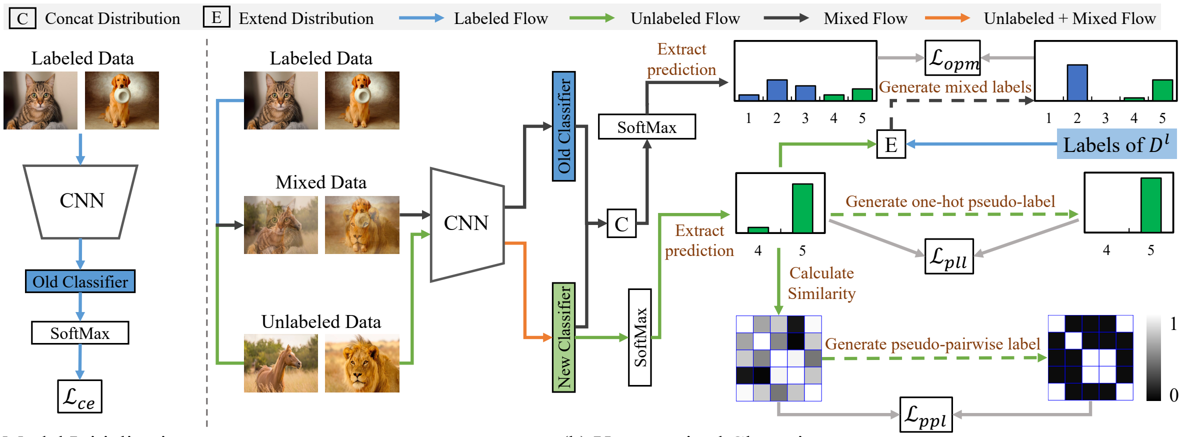
Stage 1: Model Initialization Given the labeled data D^l = $\{X^l, Y^l\}$, we are able to train the model in a supervised way. Specifically, the model is trained with the cross-entropy loss, as done in the traditional supervised classification [13, 9]. The loss function is formulated as,

$$\mathcal{L}_{ce} = -\frac{1}{n^l}\sum_{i=1}^{n^l} \log[\mathrm{SoftMax}(z_i^l)]^\top \cdot \hat{y}_i^l, \tag{1}$$

where n^l is the number of labeled training samples in a mini-batch, $z_i^l \in \mathbb{R}^{C^l}$ is the output of old classifier, and $\hat{y}_i^l \in \mathbb{R}^{C^l}$ is the one-hot label converted by y_i^l.

Stage 2: Unsupervised Clustering

Pseudo-Pair Learning. Given the model pre-trained on the labeled data, we additionally add a classifier layer of C^u new classes on the head of CNN. We then focus on the second stage, *i.e.*, unsupervised clustering in unlabeled data.



(a) Model Initialization (b) Unsupervised Clustering

Figure 2. The pipeline of the proposed method. (a) We first initialize the model on the labeled data ($\mathcal{L}_{ce}$). (b) Then, we learn the unsupervised clustering model for discovering new classes in unlabeled data, by pseudo-pair learning ($\mathcal{L}_{ppl}$), pseudo-label learning ($\mathcal{L}_{pll}$) and learning with the proposed OpenMix ($\mathcal{L}_{opm}$).

To achieve this goal, we first explore the relationship between two images for model training. Inspired by DAC [3] and DCCM [24], we argue that the relation of pairwise images is binary. In other words, each pair of images should be either of the same class or different classes. In light of this, we convert the unsupervised clustering problem to a binary classification one, aiming to distinguish whether a pair of images belong to the same class.

Similar to DAC [3] and DCCM [24], we first obtain the outputs of the new classifier for input unlabeled samples and compute their cosine similarity matrix $\mathcal{S} \in \mathbb{R}^{n^u \times n^u}$, where

$$\mathcal{S}_{i,j} = \frac{(\hat{z}^u_i)^\top . \hat{z}^u_j}{\|\hat{z}^u_i\|_2 \|\hat{z}^u_j\|_2}, \; \hat{z}^u_i = \mathrm{SoftMax}(z^u_i). \tag{2}$$

$z^u_i \in \mathbb{R}^{C^u}$ is the output of the new classifier. n^u denotes the number of unlabeled training images in a mini-batch. We then estimate the pseudo-pairwise labels $\mathcal{W}$ by setting a threshold θ_1 on $\mathcal{S}$, where

$$\mathcal{W}_{i,j} = \begin{cases} 0, & \mathcal{S}_{i,j} < \theta_1 \\ 1, & \mathcal{S}_{i,j} \geq \theta_1 \end{cases}. \tag{3}$$

By doing so, two images are defined as a positive pair if their cosine similarity is larger than θ_1, otherwise they are a negative pair. Given this pairwise supervision, we train the model with a binary cross-entropy loss, formulated as,

$$\begin{aligned} \mathcal{L}_{ppl} = &-\frac{1}{(n^u)^2} \sum_{i,j} \big(\mathcal{W}_{i,j} \log \mathcal{S}_{i,j} \\ &+ (1 - \mathcal{W}_{i,j}) \log(1 - \mathcal{S}_{i,j})\big), \forall i,j \in \{1, 2, ..., n^u\}. \end{aligned} \tag{4}$$

Pseudo-Label Learning. According to the proof given by DAC [3] and DCCM [24], the constraint $\mathcal{W}_{i,j}$ between images x^u_i and x^u_j defined in Eq. 4 can bring the following clustering property: *If the optimal solution of Eq. 4 is achieved, $\forall i, j$, $\hat{z}^u \in \mathbb{R}^{C^u}$, $\hat{z}^u_i = \hat{z}^u_j \Leftrightarrow \mathcal{W}_{i,j} = 1$, and, $\hat{z}^u_i \neq \hat{z}^u_j \Leftrightarrow \mathcal{W}_{i,j} = 0$.*
This property denotes that the predictions of the optimal new classifier, $\hat{z}^u$, are exactly C^u-diverse one-hot vectors. In other words, the unlabeled data D^u can be automatically divided into C^u partitions.

Based on this property, for unlabeled samples, we reformulate their predictions output by the new classifier to one-hot pseudo-labels, which can be used to further improve the model performance. The one-hot pseudo-label $\hat{y}^u_i$ of an unlabeled image x^u_i is generated by setting a threshold θ_2 on $\hat{z}^u_i$, where

$$\hat{y}^u_i[j] = \begin{cases} 0, & \hat{z}^u_i[j] < \theta_2 \\ 1, & \hat{z}^u_i[j] \geq \theta_2 \end{cases}. \tag{5}$$

In the pseudo-label learning, we only train the model with the unlabeled samples that are assigned with one-hot pseudo-labels, *i.e.*, $\mathrm{Max}(\hat{y}^u) = 1$. Given the one-hot pseudo-labels for unlabeled samples, we are able to train the model with cross-entropy loss, formulated as,

$$\mathcal{L}_{pll} = -\frac{1}{\hat{n}^u} \sum_i \log(\hat{z}^u_i)^\top \cdot \hat{y}^u_i, \;\; \forall i \in \{\mathrm{Max}(\hat{y}^u_i) = 1\}, \tag{6}$$

where $\hat{n}^u$ is the number of unlabeled samples that are assigned with one-hot pseudo-labels in a mini-batch.
Combination of Two Losses. By jointly considering the pseudo-pair learning and pseudo-label learning, the unsupervised clustering loss is expressed as,

$$\mathcal{L}_{uc} = \mathcal{L}_{ppl} + \lambda_1 \mathcal{L}_{pll}, \tag{7}$$

where λ_1 is the hyper-parameter that controls the importance of pseudo-label learning. To this end, we have presented our baseline for novel class discovery.

3.2. OpenMix

In the baseline presented in Section 3.1, the labeled data only play the role of model initialization. However, there is no utilization of labeled data in the second unsupervised clustering stage. In this paper, we argue that the labeled data can provide important knowledge for improving the unsupervised clustering. In this section, we propose the OpenMix for effectively leveraging the labeled data D^l during the unsupervised clustering in unlabeled data D^u. OpenMix is easy to implement. In a nutshell, during unsupervised clustering, OpenMix additionally compounds examples in two ways: 1) mix unlabeled examples with labeled samples; and 2) mix unlabeled examples with reliable anchors.

Mix with Labeled Examples. In the first way, OpenMix mixes the labeled samples with unlabeled samples, as well as their labels with pseudo-labels. Taking the prior knowledge that labeled samples and unlabeled samples belong to completely different classes, we first extend the label distributions of the labeled samples and unlabeled samples to the same size. Specifically, we concatenate $\hat{y}^l$ with a C^u-dim zeros-vector while $\hat{z}^u$ with a C^l-dim zeros-vector. The extended labels/pseudo-labels are represented by $\bar{y}^l$ for labeled samples and $\bar{y}^u$ for unlabeled samples, respectively. We then generate virtual sample with MixUp [29],

$$\begin{aligned} &\eta \sim \mathrm{Beta}(\epsilon, \epsilon), \;\; \eta^* = \mathrm{Max}(\eta, 1-\eta), \\ &m = \eta^* x^l + (1-\eta^*) x^u, \;\; v = \eta^* \bar{y}^l + (1-\eta^*) \bar{y}^u, \end{aligned} \tag{8}$$

where ϵ is a hyper-parameter and $\eta \in [0,1]$. m is the generated sample and v is the pseudo-label of m. The second constraint in Eq. 8 ensures that the generated sample m is closer to x^l than x^u. This can alleviate the negative impact caused by unreliable pseudo-labels of unlabeled samples.

As shown in Fig. 1 (b), the mixed sample has exactly true confidence in the old classes and medium confidence in the new classes. This is benefited from the prior knowledge, *i.e.*, the label of labeled sample is exactly true, and, the classes of labeled and unlabeled samples are completely different. Therefore, by mixing labeled samples with unlabeled samples through OpenMix, the pseudo-labels of mixed samples will be more reliable than that of their unlabeled counterparts. Learning with the mixed samples can help prevent the model from overfitting on unlabeled samples that are assigned with wrong pseudo-labels.

Mix with Reliable Anchors. By training with samples generated by the first way, we observe that the model keeps considerable accuracy for unlabeled samples that are predicted with high class-probabilities ($\mathrm{Max}(\hat{z}^u) \geq \theta_2$). Based on this observation, in second way, we further select the unlabeled samples that have high class-probabilities as reliable anchors. Then, we mix the anchors with unlabeled samples through OpenMix. Specifically, we perform this operation by replacing the labeled sample x^l with a reliable anchor in Eq. 8.

Loss of OpenMix. Given mixed samples $\mathcal{M}$ and their pseudo-labels $\mathcal{V}$, we apply L2-norm loss to train the model, defined as,

$$\mathcal{L}_{opm} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \frac{1}{C^l + C^u} \| v_i - \mathrm{SoftMax}(z_i^m) \|_2, \tag{9}$$

where $|\mathcal{M}|$ denotes the number of samples in $\mathcal{M}$. z_i^m indicates the concentrated outputs of the old and new classifiers. Specifically, we forward m_i to the model and extract the outputs of the old and new classifiers, which are represented as z_i^l and z_i^u, respectively. z_i^m is then obtained by concentrating z_i^l and z_i^u.

Remark. OpenMix has the following advantages: 1) By mixing labeled and unlabeled samples, the mixed pseudo-label at least contains a portion of correct labels on both known and new classes, thereby effectively eliminating incorrect pseudo-labels of unlabeled samples; 2) The joint-distribution of known and new classes injects an extra weak supervision during training, *i.e.*, labeled/unlabeled data are not of new/known classes, which can restrain the errors of pseudo-labels on new classes. The above two aspects enable our OpenMix to provide more stable and reliable training.

3.3. Overall Loss

By combining the baseline and the proposed OpenMix, the overall loss of our method is expressed as,

$$\mathcal{L}_{all} = \mathcal{L}_{uc} + \lambda_2 \mathcal{L}_{opm}, \tag{10}$$

where λ_2 balances the weight of OpenMix. During testing, we use the new classifier to predict the class of unlabeled samples.

4. Experiments

4.1. Datasets and Settings

Datasets. In this paper, we evaluate our method on three image classification benchmarks, including CIFAR-10 [12], CIFAR-100 [12] and ImageNet [4]. Following [7], we conduct the experiment on the setting where the number of classes in unlabeled data is known. **CIFAR-10** [12] includes 50,000 training images and 10,000 test images from 10 classes. Each image has a size of 32 × 32. For novel class discovery, we regard the samples of the first five classes (*i.e.,* airplane, automobile, bird, cat and deer) as labeled data while the remaining samples as unlabeled data. **CIFAR-100** [12] is similar to CIFAR-10, except that samples of CIFAR-100 are drawn from 100 classes. We regard the samples of first 80 classes as labeled data, the samples of last 10 classes as unlabeled data, and the remaining samples as validation data. **ImageNet** [4] contains 1.28 million training images from 1,000 classes. Following [22, 10, 11], we divide the ImageNet into two splits, which contain 822 and 118 classes, respectively. We use the 822-class split as the labeled set. Three 30-class subsets randomly sampled from the 118-class split are used as unlabeled sets.

Evaluation. We employ the clustering accuracy (ACC) and normalized mutual information (NMI) [18] as the metrics to evaluate the clustering performance of new classes. Both metrics range from 0 to 1. Higher scores mean better performance. For CIFAR-10 and CIFAR-100, we show the average results of 10 runs. For ImageNet, results averaged in three different subsets are reported.

4.2. Implementation Details

For a fair comparison, we follow [10, 11, 8] and use the 6-layer VGG-like architecture [17] / ResNet-18 [9] network for CIFAR-100 / {CIFAR-10, ImageNet}. For all three datasets, we pre-train the CNN and old classifier on the labeled data with common practice of supervised image classification [9]. Given the pre-trained model, we add a new classifier on the head of the CNN and train the clustering model. Specifically, we use RMSprop as the optimizer to train the model. The learning rate is kept to 0.0001 throughout the training process. We train the model for a total of 200/400/100 epochs for CIFAR-10/CIFAR-100/ImageNet. The batch sizes of unlabeled data and mixed data are both set to 64. During training, we fix the CNN and only train the new classifier at the first 60/50/100 epochs for CIFAR-10/CIFAR-100/ImageNet. Then, we train the whole model in the remaining epochs. For OpenMix, we inject the two ways of OpenMix at the 2-th epoch and 5-th epoch, respectively. For all experiments, we set $\theta_1 = 0.95$, $\theta_2 = 0.9$, $\lambda_1 = 5$, $\lambda_2 = 1000$, and $\epsilon = 1$, which can consistently achieve well performance across datasets.

4.3. Evaluation

Ablation study on baseline. In Table 1, we first investigate the two components in the baseline model, *i.e.*, pseudo-pair learning (PPL) and pseudo-label learning (PLL). For comparison, we remove one of them from the baseline model and train the model. As shown in Table 1, each component contributes to improve the performance. Among them, pseudo-pair learning is the most important to novel class discovery. Without pseudo-pair learning, the clustering accuracy is significantly reduced from 90.9% to 70.8% on CIFAR-10, and the model fails to converge on CIFAR-100 (ACC=23.9%). When removing pseudo-label learning from the baseline model, the performance will consistently be reduced on both datasets, especially on CIFAR-100. For example, on CIFAR-100, the results of the model trained without pseudo-label learning ("baseline w/o PLL") are lower than the baseline by 4% in ACC and by 0.051 in NMI. Considering these two components together achieves the best results, demonstrating their mutual benefit.

Ablation study on OpenMix. To validate the effectiveness of the proposed OpenMix, we compare OpenMix with the following variants:

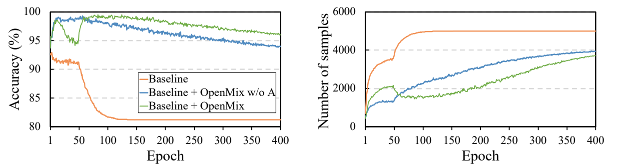
Table 1. Ablation study on CIFAR-10 and CIFAR-100. **PPL**: pseudo-pair learning, **PLL**: pseudo-label learning, **MixUp**: original MixUp [29], **Extend**: extend the label distribution, **OpenMix**: the proposed OpenMix; **L**: labeled samples, **U**: unlabeled samples, **A**: anchors selected from unlabeled samples.

Method	CIFAR-10		CIFAR-100	
	ACC	NMI	ACC	NMI
Baseline	90.9%	0.787	81.2%	0.689
Baseline w/o PPL	70.8%	0.691	23.9%	0.094
Baseline w/o PLL	90.0%	0.767	77.2%	0.638
Basel. + MixUp (U)	80.2%	0.575	78.2%	0.683
Basel. + MixUp (A)	79.4%	0.553	77.6%	0.649
Basel. + Extend (L)	90.7%	0.785	81.8%	0.709
Basel. + Extend (U)	90.8%	0.789	81.5%	0.702
Basel. + Extend (L+U)	91.4%	0.781	81.9%	0.708
Basel. + OpenMix w/o A	93.3%	0.828	84.5%	0.733
Basel. + OpenMix w/o L	75.2%	0.486	81.6%	0.690
Basel. + OpenMix	**95.3%**	**0.879**	**87.2%**	**0.754**

- MixUp (U) / MixUp (A): directly apply MixUp [29] among unlabeled samples / anchors selected from unlabeled samples.
- Extend (L) / Extend (U) / Extend (L+U): extend the label distributions of labeled samples / unlabeled samples / labeled + unlabeled samples to the size of $C^l + C^u$, and use them to train the model with cross-entropy loss.
- OpenMix w/o L: apply OpenMix without mixing labeled samples with unlabeled samples; OpenMix w/o A: apply OpenMix without mixing anchors with unlabeled samples.

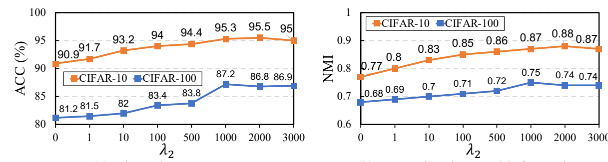
From the comparisons in Table 1, we obtain the following observations:

1. Without any modification, directly applying MixUp [29] among unlabeled samples fails to achieve an improvement. Both "Basel. + MixUp (U)" and "Basel. + MixUp (A)" reduce the results of the baseline model. The main reason is that labeled samples of new classes are not available in novel classes discovery and pseudo-labels of unlabeled samples are unreliable. Mixing samples that are assigned with unreliable pseudo-labels may still generate unreliable ones. This may lead the model to overfit on samples that are assigned with wrong pseudo-labels, thus harming the model performance.
2. Extending the label distributions to the distributions that include both old and new classes may slightly improve the performance. The improvement is benefited



(a) Classification accuracy of anchors (b) Number of anchors

Figure 3. Evolution of (a) classification accuracy of selected anchors and (b) the number of selected anchors throughout training. Experiment is conducted on CIFAR-100.



(a) Clustering accuracy (b) Normalized mutual information

Figure 4. Sensitivity to the weight of OpenMix.

from the extra weak supervision that their labels in the extended classes are correct. For example, "Basel. + Extend (L+U)" improves the clustering accuracy of the baseline from 81.2% to 81.9% for CIFAR-100.

3. Applying OpenMix between labeled samples and unlabeled samples ("Basel. + OpenMix w/o A") can consistently improve the performance of the baseline. For instance, "Basel. + OpenMix w/o A" achieves 2.4% and 3.3% improvement in clustering accuracy compared to the baseline for CIFAR-10 and CIFAR-100, respectively. In addition, applying OpenMix between labeled samples and unlabeled samples is an essential step in our method. When applying OpenMix only between anchors and unlabeled samples ("Basel. + OpenMix w/o L"), the improvement is very limited on CIFAR-100, or even negative on CIFAR-10.

4. Additionally performing OpenMix between anchors and unlabeled samples based on "Basel. + OpenMix w/o A" can further increase the performance. For example, "Basel. + OpenMix" surpasses "OpenMix w/o A" by 2% and 2.7% in clustering accuracy for CIFAR-10 and CIFAR-100, respectively. With the benefit of "Basel. + OpenMix w/o A", we are able to select reliable anchors from unlabeled samples and utilize them to dominate the mixing process for further improvement.

In conclusion, the first two observations indicate that directly using MixUp [29] or extending label distribution fails to effectively improve the performance for novel class discovery. On the other hand, the latter two observations demonstrate the effectiveness of the proposed OpenMix for novel class discovery. In addition, it should be emphasized that the proposed OpenMix is not simply implementing MixUp in novel class discovery. Instead, we explicitly consider the prior knowledge and carefully design a method for mixing examples from disjoint classes.

Investigation of selected anchors. In Fig. 3, we evaluate the classification accuracy of selected anchors and the number of selected anchors throughout the training process. It is clear that the model trained with using OpenMix among labeled data and unlabeled data ("Baseline + OpenMix w/o A") achieves much higher accuracy than the baseline model. In addition, the accuracy of "Baseline + OpenMix w/o A" can always be maintained above 95%. The full version of OpenMix, which additionally mixes selected anchors with unlabeled samples, will slightly reduce the accuracy after 20 epochs. The reduction in accuracy is mainly caused by introducing more combinations of unlabeled samples. Because the CNN is fixed at the first 50 epochs, learning with more combinations may lead the new classifier to overfit on unreliable samples. However, the accuracy of OpenMix will quickly increase after 50 epochs and will be higher than that of "Baseline + OpenMix w/o A" after 80 epochs. From Fig. 3 (b), we can observe that the numbers of selected anchors of two OpenMix-based methods are consistently lower than that of the baseline model. The above observations indicate that the proposed OpenMix can prevent the model from overfitting on wrong pseudo-labels and ensures the model to train with cleaner samples.

Impact of the weight of OpenMix. In Fig. 4, we evaluate the important hyper-parameter of our method, *i.e.*, the weight of OpenMix (λ_2). For evaluation, we keep other hyper-parameters unchanged and vary λ_2 in a range of [0, 5000]. When $\lambda_2 = 0$, our method reduces to the baseline model. When inserting OpenMix into the system ($\lambda_2 \geq 1$), the results are consistently improved in all values. Specifically, the performance first increases with λ_2 and becomes stable when $\lambda_2 \geq 1000$. The best results are produced when λ_2 in the range of [1000, 3000]. This indicates that our method is insensitive to the changing of λ_2 in a wide range.

4.4. Comparison with State-of-The-Art

We compare the proposed method with the state-of-the-art in novel class discovery, including: K-means [14], KCL [10], MCL [11], DTC [8] and RS [7]. For K-means [14], we train the model on the labeled data and extract the last layer of the CNN as the features of unlabeled samples. Then, we directly perform K-means on unlabeled data to obtain clustering results. Comparisons are shown in Table 2. Our baseline achieves very competitive clustering performance compared with the state of the art. The baseline is higher than DTC [8] on CIFAR-10 and CIFAR-100, and slightly lower than DTC [8] on ImageNet. In addition, it is clear that our method ("Baseline+OpenMix") outperforms state-of-the-art methods by a large margin. Specifically, our approach achieves **95.3% for CIFAR-10, 90.1% for CIFAR-**

Table 2. Comparison with state-of-the-art methods on CIFAR-10, CIFAR-100 and ImageNet for novel class discovery. Note that, RS [7] did not evaluate the NMI metric and did not provide results of 10-class setting on CIFAR-100.

Method	Venue	CIFAR-10		CIFAR-100		ImageNet	
		ACC	NMI	ACC	NMI	ACC	NMI
K-means [14]	Classic	65.5%	0.422	66.2%	0.555	71.9%	0.713
KCL [10]	ICLR18	66.5%	0.438	27.4%	0.151	73.8%	0.750
MCL [11]	ICLR19	64.2%	0.398	32.7%	0.202	74.4%	0.762
DTC [8]	ICCV19	87.5%	0.735	72.8%	0.634	78.3%	0.791
RS [7]	ICLR20	91.7%	-	-	-	82.5%	-
Baseline	This	90.9%	0.787	81.2%	0.689	77.1%	0.784
Ours	Work	**95.3%**	**0.879**	**87.2%**	**0.754**	**85.7%**	**0.827**

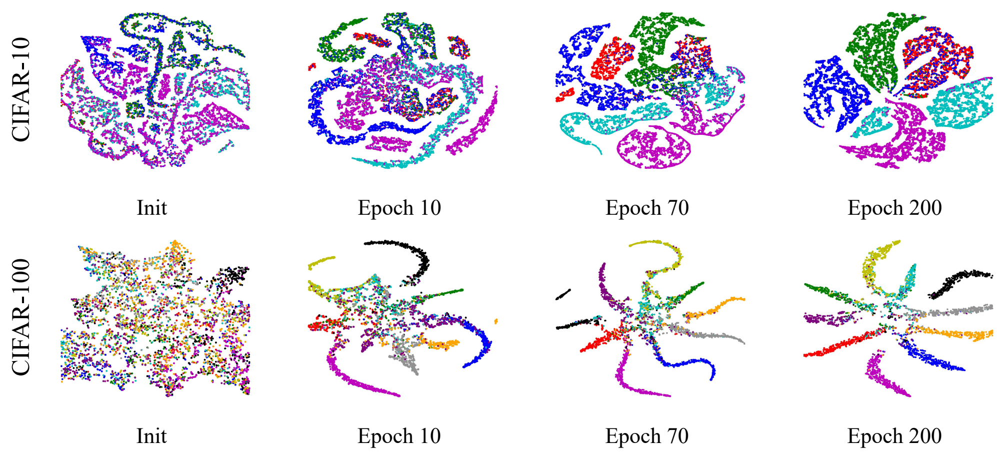


Figure 5. T-SNE [21] visualization of unlabeled samples on CIFAR-10 and CIFAR-100. Results in different stages of our method are shown. Different colors represent different classes. Our method progressively separates the unlabeled samples into discriminative clusters.

100 and 85.7% for ImageNet in clustering accuracy. Both KCL [10] and MCL [11] use pairwise similarity for clustering learning. However, these two method fail to produce competitive performance. For example, KCL [10] and MCL [11] have similar resutls to K-means on CIFAR-10 and ImageNet, but are largely inferior to K-means on CIFAR-100. Our method is significantly superior to KCL [10] and MCL [11]. Compared to DTC [8], our method surpasses it in all three datasets, especially in CIFAR-100. RS [7] is the latest method, which also uses the labeled data during unsupervised clustering. However, RS mainly focuses on using labeled data to maintain the accuracy in old classes. Compared to RS [7], our method outperforms it by 3.6% and 3.2% in clustering accuracy for CIFAR-10 and ImageNet, respectively, indicating that our method establishes the new state of the art result.

4.5. Visualization

To better reflect the effectiveness of the proposed method, we visualize the distributions of unlabeled samples in different stages. Specifically, we extract the outputs of the new classifier as features of samples and map them into 2-D vectors by t-SNE [21]. Visualization results are shown in Fig. 5. In initialization, unlabeled samples are scattered. Through the training of our method, the samples of the same class are progressively grouped together, thereby enabling the new classifier to accurately recognize new classes in unlabeled data.

5. Conclusions

This work studies the problem of discovering novel classes in unlabeled data given labeled data of disjoint classes. To address this problem, we focus on effectively incorporating the labeled data into the step of unsupervised clustering in unlabeled data. To achieve this goal, we present OpenMix to dynamically incorporate labeled samples of known classes and unlabeled samples of novel classes as well as their labels and pseudo-labels. OpenMix can generate joint-class samples with reliable pseudo-labels and diverse smooth samples of new classes. Learning with these generated samples helps to improve the model performance of new class recognition. Experiments conducted on three image classification benchmarks demonstrate that OpenMix can consistently improve the performance of a competitive baseline, enabling us to achieve state-of-the-art results.

References

[1] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *Proc. ICLR*, 2020. 2, 3

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Proc. NeurIPS*, 2019. 1, 2, 3

[3] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proc. ICCV*, 2017. 4

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009. 1, 5

[5] Kamran Ghasedi Dizaji, Amirhossein Herandi, Cheng Deng, Weidong Cai, and Heng Huang. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *Proc. ICCV*, 2017. 3

[6] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proc. NeurIPS*, 2018. 1

[7] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *Proc. ICLR*, 2020. 1, 3, 5, 7, 8

[8] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proc. ICCV*, 2019. 1, 3, 6, 7, 8

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 1, 3, 6

[10] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *Proc. ICLR*, 2018. 1, 3, 5, 6, 7, 8

[11] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *Proc. ICLR*, 2019. 1, 3, 5, 6, 7, 8

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NeurIPS*, 2012. 3

[14] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proc. BSMSP*, 1967. 3, 7, 8

[15] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Proc. NeurIPS*, 2018. 1, 3

[16] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *TKDE*, 2009. 3

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2015. 1, 6

[18] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *JMLR*, 2002. 6

[19] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *Proc. ICANN*, 2018. 3

[20] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. NeurIPS*, 2017. 3

[21] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *JMLR*, 2014. 8

[22] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Proc. NeurIPS*, 2016. 5

[23] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 2016. 3

[24] Jianlong Wu, Keyu Long, Fei Wang, Chen Qian, Cheng Li, Zhouchen Lin, and Hongbin Zha. Deep comprehensive correlation mining for image clustering. In *Proc. ICCV*, 2019. 4

[25] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proc. ICML*, 2016. 3

[26] Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proc. ICML*, 2017. 3

[27] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proc. ICCV*, 2019. 2, 3

[28] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Proc. NeurIPS*, 2005. 3

[29] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. ICLR*, 2018. 2, 3, 5, 6, 7