Novel Class Discovery in Chest X-rays via Paired Images and Text

Jiaying Zhou[1, 2], **Yang Liu**[3], **Qingchao Chen**[1, 2, 4 *]

[1]National Institute of Health Data Science, Peking University, Beijing, China
[2]Institute of Medical Technology, Peking University Health Science Center, Beijing, China
[3]Wangxuan Institute of Computer Technology, Peking University, Beijing, China
[4] National Key Laboratory of General Artificial Intelligence, Beijing, China
{zhoujiaying, qingchao.chen, yangliu}@pku.edu.cn

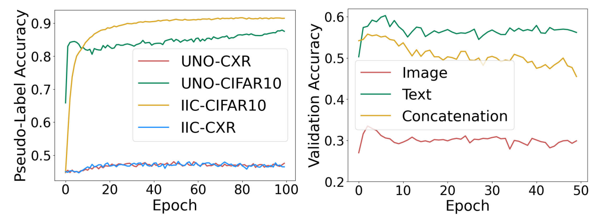
Abstract

Novel class discover(NCD) aims to identify new classes undefined during model training phase with the help of knowledge of known classes. Many methods have been proposed and notably boosted performance of NCD in natural images. However, there has been no work done in discovering new classes based on medical images and disease categories, which is crucial for understanding and diagnosing specific diseases. Moreover, most of the existing methods only utilize information from image modality and use labels as the only supervisory information. In this paper, we propose a multi-modal novel class discovery method based on paired images and text, inspired by the low classification accuracy of chest X-ray images and the relatively higher accuracy of the paired text. Specifically, we first pretrain the image encoder and text encoder with multi-modal contrastive learning on the entire dataset and then we generate pseudo-labels separately on the image branch and text branch. We utilize intra-modal consistency to assess the quality of pseudo-labels and adjust the weights of the pseudo-labels from both branches to generate the ultimate pseudo-labels for training. Experiments on eight subset splits of MIMIC-CXR-JPG dataset show that our method improves the clustering performance of unlabeled classes by about 10% on average compared to state-of-the-art methods. Code is available at: https://github.com/zzzzzzzzjy/MMNCD-main.

Introduction

The success of deep learning based classification methods greatly depends on labeled data. However, it is difficult to gather high-quality labeled data, especially for medical data. And in real-world scenarios, it is almost impossible to collect labeled data for all classes because of missing definitions, vague categories, infinite categories, etc. To address this problem, a new paradigm called *novel class discovery(NCD)* has been proposed and gained significant attention due to its potential applications in various domains, such as surveillance, medical image analysis, and anomaly detection.

Given a labeled set, the goal of NCD is to discover undefined categories in the unlabeled set, which distinguishes it



(a) Pseudo-label accuracy. (b) Validation set accuracy.

Figure 1: (a)Pseudo-label accuracy of different methods and different datasets. (b)Validation set accuracy of unlabeled classes in *SET1* under supervision.

from semi-supervised learning. Most of the existing methods start with supervised pretraining on the labeled set, while others adopt self-supervised pretraining on the whole dataset. Two-stage methods then employ learned similarity prediction networks or feature extraction networks to classify the unlabeled set through clustering or (pairwise) pseudo-labeling(Hsu, Lv, and Kira 2017; Hsu et al. 2019; Han, Vedaldi, and Zisserman 2019). Some one-stage methods tune the feature representation while classifying/clustering by using different objective functions for labeled set and unlabeled set(Han et al. 2020; Zhong et al. 2021a; Han et al. 2021). Others unify the objective function as cross-entropy loss by assigning pseudo-labels to unlabeled samples(Fini et al. 2021; Li et al. 2022a; Yang et al. 2022). In this paper, we mainly focus on one-stage methods that unify the objective function through pseudo-labels.

Despite abundant research on the topic of NCD, most of these works are conducted on benchmark datasets of natural images. The usability and effectiveness of the NCD methods on large-scale medical image datasets are not yet known. NCD for medical images is of great importance for disease diagnosis and precision medicine because of its ability to uncover new disease types or unknown disease subtypes. In this paper, we focus on NCD in Chest X-Ray (CXR) images that are widely used in medical diagnostics for detecting lung diseases, tumors, and other abnormalities.

The success of pseudo-label-based one-stage methods relies heavily on the quality of the pseudo-labels(Lee et al. 2013). Existing methods (Fini et al. 2021; Li et al. 2022a)

*Corresponding author.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

formulate pseudo-label assignment as an optimal transport problem that can be solved by Sinkhorn-Knopp algorithm(Cuturi 2013). Also, they improve the reliability of pseudo-labels by constraining the consistency of the predictions of the image and its augmented view. *However,* the pseudo-label assignment accuracy of these methods on the CXR dataset is much lower than that on the natural image datasets with a similar number of classes, as shown in Fig.1. We speculate that this may be due to the highly similar appearances of CXR images from different anatomical structures or classes. With these pseudo-labels, the mainstream methods do not achieve good results on the CXR dataset. Meanwhile, we find that the classification accuracy of CXR images under supervision is also not very good, at least not as good as text classification, as shown in Fig.1. Inspired by this, we hold the hypothesis that text can help generate better pseudo-labels in some cases and aim to investigate the usage of the descriptive text paired with CXR images to improve performance. We would like to solve the following research questions: when does one modality provide better pseudo-labels than another and how to boost the image NCD performance utilizing the advantageous text information in the training process.

To answer the two questions and innovate text-augmented CXR NCD, it still entails three technical challenges: given the heterogeneous cross-modal semantic gap, (1) how to quantify the quality of pseudo-labels calculated by visual and text modalities; (2) how to transfer useful information between two modalities to improve CXR NCD performance through text; (3) the unavailable text at the test time in the real-world deployments.

Constrained by unavailable text at the test time, instead of multi-modal feature fusion, we propose to use a two-branch network, encoding the visual and text features separately. To quantify the pseudo-label quality of two modalities, we propose a novel measurement – the consistency between *semantic feature structure* and the *pseudo-label structure* in each modality. The quality of the pseudo-labels in NCD relies on two aspects: (1) how much the abstract features can distinguish the unknown classes; (2) how the semantic patterns in the local feature embeddings are able to identify and comprehend the relationship between the known and the novel classes. Pseudo-labels or the features close to them may represent abstract and categorical information but tend to *lose* local visual semantic features. Therefore, to bring advantages from both worlds, i.e. the categorical and local semantic features, we propose to utilize their structural consistency for pseudo-label quality identification. By comparing the consistencies in both modalities, we can identify when to transfer information from one modality to another.

To integrating effective information from both modalities and reduce the cross-modal semantic gap, we propose to synthesize pseudo-labels that guide both visual and text NCD. The synthetic pseudo-labels are the linear combination of the visual and text ones, using the quantified consistency scores respectively. The potential reason is that visual and text features are heterogeneous, although the cross-modal alignment losses are performed(Liang et al. 2022), direct transfer/distillation or feature alignment may result in degraded NCD performance.

Our contributions can be summarized as follows:

- We explore NCD in CXR images and propose a method based on paired images and text.
- We introduce intra-modal consistency as a basis for pseudo-label quality measuring and weighting.
- Based on the MIMIC-CXR-JPG dataset, we set up two benchmarks which have the same known classes and different new classes. We evaluate the proposed methods on eight data splits of the two benchmarks and demonstrate significant performance improvements over the state-of-the-art methods.

Related Work

Novel Class Discovery

Novel class discovery(NCD) aims to discover new classes in an unlabeled dataset given different but related labeled classes. Existing methods can be divided into two categories: two-stage methods and one-stage methods.

The pioneering works of NCD are two-stage methods, including KCL(Hsu, Lv, and Kira 2017), MCL(Hsu et al. 2019) and DTC(Han, Vedaldi, and Zisserman 2019). The KCL and MCL utilize similarity prediction networks to generate pairwise pseudo-labels and leverage the clustering models to classify unlabeled data. The two stages adopt different objective functions. DTC first trains a model with supervised learning on the labeled set and then discovers novel visual categories using DEC(Xie, Girshick, and Farhadi 2016). MM/MP(Chi et al. 2021) trains a group of classifiers on the labeled set and fine-tunes classifiers on the unlabeled set.

Compared to two-stage methods, one-stage methods received more attention in the field recently. One-stage methods use both labeled data and unlabeled data simultaneously at some point in the optimization process. RS/AutoNovel(Han et al. 2020, 2021) may be the first work among the one-stage methods. It uses the pairwise similarity obtained by ranking statistics as supervision to discover novel classes. Follow-up work DualRS(Zhao and Han 2021) expands this method to a two-branch framework focusing on both local and global features. Afterward, NCL(Zhong et al. 2021a) further boosts the performance by leveraging the framework of contrastive learning. In addition, OpenMix(Zhong et al. 2021b) uses MixUp(Zhang et al. 2017) to generate more robust pseudo-labels for the unlabeled data.

Other one-stage methods eliminate the use of pairwise pseudo-labels and directly assign pseudo-labels to unlabeled samples. UNO(Fini et al. 2021) may be the first of these works. UNO unifies the training objective by using a multi-view self-labeling strategy to generate pseudo-labels that can be treated homogeneously with ground truth labels. Based on UNO, IIC(Li et al. 2022a) models both inter-class and intra-class constraints based on the symmetric Kullback-Leibler divergence. ComEx(Yang et al. 2022) focuses on the generalized setting of NCD(GNCD) and classifies the data with two complementary groups of classifiers with global-to-local and local-to-local regularization to strengthen pseudo-labels. Similar to the ComEx, some work focuses on generalized class discovery(GCD). Among them, CLIP-GCD is the first work to combine multi-modal (image and text) models in GCD. CLIP-GCD proposes a retrieval-based mechanism that leverages CLIP's aligned visual-language representations.

Different from the previous works, we propose to solve a novel task, text-augmented NCD in medical image analysis. Our work focuses on quantifying the quality of pseudo-labels from both modalities and integrating effective information from both text and image to boost CXR NCD performance. We did not adopt the solution to retrieve text annotations for medical images from the text corpus due to unavailable cross-modal pre-trained models like CLIP for medical images. In addition, the medical image exhibits unique challenges for NCD, i.e., high semantic similarity in the anatomical structures. To our best knowledge, this is the first work that tackle the NCD in medical image analysis using both image and text.

Pseudo-Labeling in Semi-Supervised Learning

Our work is related to part of the work on semi-supervised learning involving pseudo-labeling. Among them, MixMatch(Berthelot et al. 2019b) averages and sharpens the predictions of multiple strongly augmented views as pseudo-labels. ReMixMatch(Berthelot et al. 2019a) proposes to generate the pseudo-labels with weakly augmented views and align the pseudo-label distribution with the marginal distribution of ground-truth labels. Instead of using all pseudo-labels, FixMatch(Sohn et al. 2020) retains only those with high confidence. SoftMatch(Chen et al. 2023) overcomes the trade-off between quantity and quality of pseudo-labels with truncated Gaussian weighting function and uniform alignment.

Different from these methods, we do not employ data augmentation and rely solely on the images and paired text to generate pseudo-labels. Our contribution lies in not only the quality quantification of pseudo-labels but also in proposing a new strategy to integrate two modality information for joint learning.

Method

Overall

Problem Formulation: Similar to the image-only NCD setting, our training data are split into a labeled set and an unlabeled set. The labeled set $D^l = \{(\mathbf{v}_1^l, \mathbf{t}_1^l, y_1^l), ..., (\mathbf{v}_N^l, \mathbf{t}_N^l, y_N^l)\}$ contains paired image and text $(\mathbf{v}_i^l, \mathbf{t}_i^l)$ with corresponding label y_i^l from C^l classes. The unlabeled set $D^u = \{(\mathbf{v}_1^u, \mathbf{t}_1^u), ..., (\mathbf{v}_M^u, \mathbf{t}_M^u)\}$ contains unlabeled paired image and text $(\mathbf{v}_i^u, \mathbf{t}_i^u)$ from C^u classes, where C^u is known as a prior. The set of C^l labeled classes is disjoint with the set of C^u unlabeled classes. The purpose of NCD is to discover C^u clusters in the unlabeled set. Following the UNO(Fini et al. 2021), we formulate this problem as learning a mapping from the sample to the complete-label set $\mathcal{Y} = \{1, ..., C^l, C^l + 1, ..., C^l + C^u\}$. To generalize the method to real-world scenarios, we assume that *the text is not available at the test time*.

Architecture: We propose a method based on paired images and text, using the intra-modal structural consistency to generate and adjust pseudo-labels. Our network architecture is shown in Fig.2, where it consists of two branches: the image branch and the text one.

Given a CXR image $\mathbf{v}$, the semantic embedding $z_v \in \mathcal{R}^k$ is firstly obtained via the visual encoder E_v and the projection head $Proj_v$, i.e. $\boldsymbol{z}_v = Proj_v(E_v(\mathbf{v}))$. Then, the two visual classification heads, labeled head h_v and unlabeled head g_v, predict their categorical contents(logits), $\boldsymbol{l}_{h_v}$ and $\boldsymbol{l}_{g_v}$, leveraging the semantic embeddings. Finally, we concatenate both the logits from h_v and g_v as follows: $\boldsymbol{l}_v = [\boldsymbol{l}_{h_v}, \boldsymbol{l}_{g_v}]$ and get the probability distribution $\boldsymbol{p}_v = \sigma(\boldsymbol{l}_v/\tau)$ via a softmax layer σ, with τ as the temperature parameter. Considering the unavailable text at test time, we utilize a parallel text branch adopting the same architecture. To be specific, the text semantic embedding $\boldsymbol{z}_t \in \mathcal{R}^k$ is obtained from text encoder E_t and projection head $Proj_t$. The text classification logits are predicted as $\boldsymbol{l}_t = [\boldsymbol{l}_{h_t}, \boldsymbol{l}_{g_t}]$ via the text classification heads h_t(labeled head), g_t(unlabeled head), so as the probability predictions $\boldsymbol{p}_t = \sigma(\boldsymbol{l}_t/\tau)$.

Given our setup of CXR NCD using text data, it is essential to quantify the quality of pseudo-labels from both the visual and text modalities. As it is challenging to quantify the heterogeneous feature structures, we propose a solution to calculate the consistency index between the semantic feature structure and the pseudo-label structure. A higher consistency score may indicate better pseudo-label quality, as the score compromises the capabilities of distinguishing the unknown classes and capturing the relationship between the known and the unknown classes. The cross-modal comparison in the space of our proposed consistency alleviates the cross-modal embedding gap in the processing. More technical details are in the following section.

Once the quality of pseudo-labels from both image and text branches is quantified, we propose to generate synthetic pseudo-labels as supervision of both branches. The synthetic pseudo-labels are integrated by weighting the pseudo-labels from images and text using the *calculated consistency scores*. Our design is able to transfer effective information between image and text branches by scheduling based on the quantified quality of each modality. Note this procedure effectively reduce the inter-modal gap as well.

Pseudo-Labeling with Estimated Prior Distribution

Pseudo-labeling is a key step of NCD. Following prior works(Fini et al. 2021; Li et al. 2022a; Yang et al. 2022), we re-formulate the clustering problem into an optimal transport(OT) problem that finds the optimal transportation between the sample distribution and the class distribution. Formally, given logits $\boldsymbol{L}_v$ of a batch of data with batch size B, we select the logits of all the unlabeled samples: $\boldsymbol{L}_v^u = [\boldsymbol{l}_{g_v}^1, ..., \boldsymbol{l}_{g_v}^{B^u}]$, our goal is to assign pseudo-labels $\hat{\mathbf{Y}}_v^u = [\hat{\boldsymbol{y}}_v^1, ..., \hat{\boldsymbol{y}}_v^{B^u}]$, where the rows of $\boldsymbol{L}_v$ represent logits while the rows of $\hat{\mathbf{Y}}_v$ represents the pseudo-labels for unlabeled samples, and B^u is the number of unlabeled samples in the batch. The problem can be solved by the Sinkhorn-Knopp algorithm as follows:

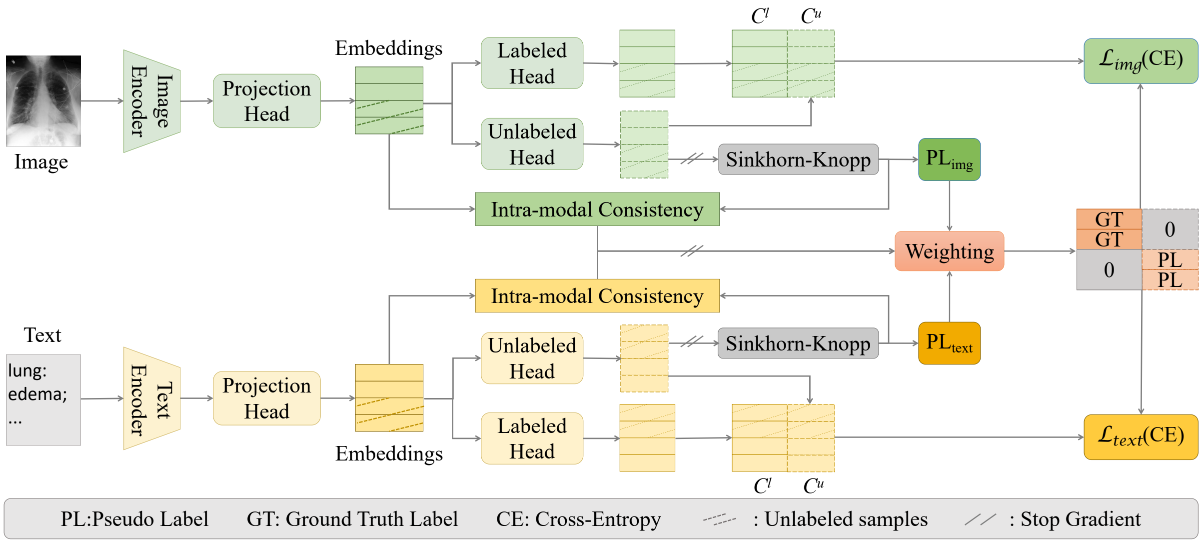


Figure 2: Overview of the proposed architecture. We present *image* branch in *green* and *text* branch in *yellow*. We generate pseudo-labels on each branch and use intra-modal consistency for weighting to generate final pseudo-labels(in pink). Cross-entropy loss is calculated based on logits, pseudo-labels and ground-truth labels. When training, the parameters of both branches are updated simultaneously; when testing, the clustering accuracy is calculated based only on the pseudo-labels assigned by the image branch.

$$\hat{\mathbf{Y}}_v^u = \max_{\mathbf{Y}_v^u \in \Gamma_v} Tr(\mathbf{Y}_v^u \boldsymbol{L}_v^u) + \epsilon H(\mathbf{Y}_v^u) \qquad (1)$$

where $\epsilon > 0$ is a hyper-parameter, H represents the entropy function used to constrain the pseudo-labels and Tr is the trace operation. Γ_v is the transportation prototype defined as:

$$\Gamma_v = \{\mathbf{Y}_v^u \in \mathcal{R}_+^{C^u \times B^u} | \mathbf{Y}_v^u \mathbf{1}_{B^u} = \boldsymbol{p}_v, \mathbf{Y}_v^{u\top} \mathbf{1}_{C^u} = \frac{1}{B^u} \mathbf{1}_{B^u}\} \qquad (2)$$

A common setting is to assume that $\boldsymbol{p}_v = \frac{1}{C^u}\mathbf{1}_{C^u}$, i.e. to distribute the samples uniformly across classes. *However*, in our CXR NCD scenario, data are not always uniformly distributed, e.g. the number of samples for rare diseases is always smaller than the number of samples for common diseases. Instead of assuming that the unlabeled samples are uniformly distributed across classes, we follow BYOP(Yang et al. 2023) to estimate and iteratively update the novel class distribution prior $\boldsymbol{p}_v$ in the OT procedure and optimization. Same as the operation in the image branch, we can also obtain pseudo-labels $\hat{\mathbf{Y}}_t^u = [\hat{\boldsymbol{y}}_t^1, ..., \hat{\boldsymbol{y}}_t^{B^u}]$ in the text branch.

Pseudo-Label Quality Estimation and Synthetic Pseudo-Label Generation

Due to the low pseudo-label accuracy generated by image branch discussed in the introduction, we aim to improve the pseudo-label quality by leveraging the text information and interaction between the two branches. A straightforward idea is to take the average of the pseudo-labels of both branches so that all the high-quality pseudo-labels are retained to some extent. *However,* the unreliable pseudo-labels are equally retained and transferred. It tends to propagate and accumulate erroneous information and prevent taking advantage of high-quality pseudo-labels. Therefore, we propose a novel and robust measurement to quantify the pseudo-label quality from both modalities. Then, leveraging this quantification score, advantageous and complementary information from both modalities can be scheduled and utilized to boost the NCD performance.

Before introducing the details of our proposal, it seems essential to define the spectrum of pseudo-label quality in the NCD problem, especially without the ground-truth labels in novel classes. Different from the semi-supervised problem, in NCD, the capability to extend and explore new classes leverages not only the abstract information (e.g., logits) but also the local visual semantics that are shared among unknown and known categories(Sun et al. 2023). The mainstream theory(Li et al. 2022b) demonstrated that the NCD performance depends on how similar/shareable local semantic attributes are across known and novel classes. Therefore, there are at least two aspects to quantify the quality of the pseudo-labels: (1) how much the abstract features(logits/pseudo-labels) can distinguish the unknown classes; (2) how the semantic patterns in the local feature embeddings are able to identify and comprehend the relationship between the known and the novel classes.

To bring advantages from both worlds, i.e. the abstract and local feature embeddings, we propose to utilize their structural consistency for pseudo-label quality identification. To put it simply, we propose a new hypothesis that if samples with more similar local semantic attributes/features, their structural similarity of pseudo-labels should be maintained. Therefore, the higher the consistency between the local semantic structure and pseudo-label structure is, the better the pseudo-label quality is. And vice versa.

Specifically, let $\mathbf{Z}_v^u = [\boldsymbol{z}_v^1, ..., \boldsymbol{z}_v^{B^u}]^\top \in \mathcal{R}^{B^u \times k}$ be the embeddings of unlabeled images in the batch, the local semantic similarity can be calculated as:

$$\boldsymbol{Sim}_{emb_v} = \mathbf{Z}_v^u \mathbf{Z}_v^{u\top} \tag{3}$$

where $\boldsymbol{Sim}_{emb_v}^{i,j} = \boldsymbol{z}_v^i \cdot \boldsymbol{z}_v^j$ represents the similarity between i_{th} image embedding and j_{th} image embedding.

Similarly, given pseudo-labels $\hat{\mathbf{Y}}_v^u = [\hat{\boldsymbol{y}}_v^1, ..., \hat{\boldsymbol{y}}_v^{B^u}]^\top \in \mathcal{R}^{B^u \times C^u}$ of unlabeled images, we can get similarity matrix of pseudo-labels as follow:

$$\boldsymbol{Sim}_{pl_v} = \hat{\mathbf{Y}}_v^u \hat{\mathbf{Y}}_v^{u\top} \tag{4}$$

where $\boldsymbol{Sim}_{pl_v}^{i,j} = \hat{\boldsymbol{y}}_v^i \cdot \hat{\boldsymbol{y}}_v^j$ represents the similarity between i_{th} image's pseudo-label and j_{th} image's pseudo-label.

For the i_{th} image, we propose to use JS-divergence to measure the consistency Con_v^i of embedding similarity and pseudo-label similarity, where

$$Con_v^i = max(m, 1 - \lambda D_{JS}(\boldsymbol{Sim}_{pl_v}^i || \boldsymbol{Sim}_{emb_v}^i)) \tag{5}$$

and m is a threshold that prevents consistency from being 0.

Following the above steps, we can calculate the embedding similarity $\boldsymbol{Sim}_{emb_t}$ and pseudo-label similarity $\boldsymbol{Sim}_{pl_t}$ of the text modality as well. And for the i_{th} text data, we use JS-divergence to measure the consistency Con_t^i of embedding similarity and pseudo-label similarity, where

$$Con_t^i = max(m, 1 - \lambda D_{JS}(\boldsymbol{Sim}_{pl_t}^i || \boldsymbol{Sim}_{emb_t}^i)) \tag{6}$$

After calculating the intra-modal consistency, we propose to generate synthetic pseudo-labels that guide the learning process for both modalities. Specifically, using the quality indexes as pseudo-label weights, the synthetic pseudo-labels of the i_{th} sample pair can be expressed as:

$$\boldsymbol{pl}^i = \frac{Con_v^i}{Con_v^i + Con_t^i} \hat{\boldsymbol{y}}_v^i + \frac{Con_t^i}{Con_v^i + Con_t^i} \hat{\boldsymbol{y}}_t^i \tag{7}$$

We train both branches using synthetic pseudo-labels. Following UNO(Fini et al. 2021), for samples from the labeled set, we zero-pad $\boldsymbol{y}^l$, i.e. $\boldsymbol{y} = [\boldsymbol{y}^l, \mathbf{0}_{C^u}]$; for samples from unlabeled set, we zero-pad $\boldsymbol{pl}^u$, i.e. $\boldsymbol{y} = [\mathbf{0}_{C^l}, \boldsymbol{pl}^u]$. Then we can train the whole network using standard cross-entropy:

$$\mathcal{L}_{img} = -- \frac{1}{B} \sum_{b=1}^{B} \sum_{c=1}^{C} \boldsymbol{y}^b(c) log(\boldsymbol{p}_t^b(c)) \tag{8}$$

$$\mathcal{L}_{text} = -\frac{1}{B} \sum_{b=1}^{B} \sum_{c=1}^{C} \boldsymbol{y}^b(c) log(\boldsymbol{p}_t^b(c)) \tag{9}$$

$$\mathcal{L}_{cls} = \mathcal{L}_{img} + \mathcal{L}_{text} \tag{10}$$

where $C = C^u + C^l$, $\boldsymbol{y}^b(c)$ is the c-th element of the label $\boldsymbol{y}^b$ of the b-th sample in a batch, $\boldsymbol{p}_v^b(c)$ is the c-th element of b-th image's prediction $\boldsymbol{p}_v^b$, $\boldsymbol{p}_t^b(c)$ is the c-th element of b-th text's prediction $\boldsymbol{p}_t^b$.

Experiment

Experiment Setup

Datasets

MIMIC-CXR-JPG Dataset(Johnson et al. 2019b) This dataset contains 377,110 chest X-ray images from 65,379 patients. Each image is provided with 14 labels derived from two natural language processing tools applied to the corresponding free-text radiology reports. In our experiment, we only focus on investigating images from the frontal view. Based on the relationship between classes and the number of samples in classes, 11 classes were selected. We divided these classes into three groups, one group as labeled classes and the remaining two groups as unlabeled classes. The labeled classes include *No Finding*, *Atelectasis*, *Lung Opacity* and *Edema*. The first group of unlabeled classes are all lung diseases, including *Consolidation*, *Pneumonia*, *Pneumothorax*. The second group of unlabeled classes are diseases that occur in other anatomical structures, including *Cardiomegaly*, *Enlarged Cardiomediastinum*, *Fracture*, *Pleural Effusion*. We refer to the combination of the labeled classes and the first group of unlabeled classes as *SET1* and the combination of the labeled classes and the second group of unlabeled classes as *SET2*. We adjust the number of samples for each class and obtain eight different dataset splits. The details are shown in the Table 1.

Chest ImaGenome Dataset(Wu et al. 2021) Chest ImaGenome dataset is automatically constructed from the MIMIC-CXR dataset(Johnson et al. 2019a). This dataset uses a rule-based text-analysis pipeline to correlate anatomies with various CXR attributes extracted from text reports. To reduce noise and prevent label leakage, we filter the attributes and formalize the text report into the form of "Anatomy-1: Attribute-1,...,Anatomy-k:Attribute-j", where k and j mean the j-th attribute of k-th anatomy. All our experiments are conducted based on formalized text.

Evaluation Metrics Adhering to the evaluation protocols employed in existing studies(Fini et al. 2021; Li et al. 2022a), our experiments are also conducted under both task-aware and task-agnostic protocols. Under the task-aware protocol, we are aware of if the paired image and text originate from the labeled set or the unlabeled set. Conversely, under the task-agnostic protocol, such information is unavailable. We use the average clustering accuracy to evaluate the performance of our method on unlabeled sets. It is defined as:

$$Cluster\_Acc = \max_{perm \in P} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{y_i = perm(pl_i)\} \tag{11}$$

where y_i and pl_i represent the ground-truth label and pseudo-label of sample $(\mathbf{v}_i, \mathbf{t}_i)$. P indicates the set of all permutations. The optimal permutation can be calculated by the Hungarian algorithm(Kuhn 2005).

Implementation Details We use ResNet-50(He et al. 2016) as the image encoder and BioClinicalBERT(Alsentzer et al. 2019) as the text encoder. We train our model in two stages. First, we finetune the encoders following GLoRIA(Huang et al. 2021) with all training data. Then we conduct novel class discovery on our network with 200 epochs. All experiments are conducted with a fixed batch size of 128.

Size	SET1			SET2		
	Labeled	Unlabeled	Sum	Labeled	Unlabeled	Sum
I	4000	3000	7000	4000	4000	8000
II	12000	3000	15000	12000	4000	16000
III	12000	*6000*	18000	12000	*8200*	20200
IV	*21000*	*6000*	27000	*21000*	*8200*	29200

Table 1: Train-set size of the splits used in our benchmark. *6000* means the number of samples for the three unlabeled classes of SET1 is 1000/2500/2500 respectively. *8200* means the number of samples for the four unlabeled classes of SET2 is 3000/1000/1200/3000 respectively. *21000* means the number of samples for the four labeled classes is 6000/3000/6000/6000 respectively. Unspecified indicates that the number of samples is the same for each class.

Following the literature(Fini et al. 2021; Li et al. 2022a; Yang et al. 2022), we use multi-head clustering and over-clustering to boost the clustering performance. The methods under comparison all use the same setup.

Comparison with State-of-the-Arts

We compare our method with the current state-of-the-art methods, including AutoNovel(Han et al. 2021), NCL(Zhong et al. 2021a), UNO(Fini et al. 2021), IIC(Li et al. 2022a) and ComEx(Yang et al. 2022) besides K-Means(McQueen 1967). We also combine our method with pseudo-label assignment based on novel class distribution prior estimation in BYOP(Yang et al. 2023) for the unbalanced data. We report the experimental results under the task-aware and task-agnostic protocol in Table 2 and Table 3 respectively.

In Table 2, we report the average clustering accuracy on the unlabeled test sets using the task-aware protocol. As we can see, our method achieves the best results on all data splits, with a substantial performance improvement over other methods. The results demonstrate the significant gain of introducing text to NCD in CXR images. Note that when using the method of novel class distribution prior estimation in BYOP(Yang et al. 2023) to guide the pseudo-label assignment, our method shows better clustering performance, although the imbalance of our data splits is relatively low. We have also tried a multi-modal knowledge distillation for NCD(*UNO_MMKD* in Table 2). Specifically, we first train the text encoder and classification head. Then, the parameters of the text branch are frozen, and the structure of the image branch is kept the same as UNO. We conduct knowledge distillation at the logits end, but the performance improvement is not obvious and the advantages of multi-modal are not utilized.

We also report the classification accuracy of labeled test sets and the average clustering accuracy of unlabeled test sets of *SET1* using the task-agnostic protocol in Table 3. In this setting, our method is significantly better at classifying labeled sets than other methods, improving classification performance by about 8% on average across four splits. Although the performance advantage of our method for unlabeled clustering is not as pronounced, we still achieve about 5% performance improvement on split II and split IV, and the clustering performance of our method is close to the optimal on split I and split III. Results on *SET2* are placed in the supplementary material due to space issues.

Analysis

Ablation Study

Intra-modal Consistency Weighting v.s. Averaging In this paper, we propose a novel pseudo-label quality quantification index and use it to weigh the pseudo-labels of the two branches as the final pseudo-label. We hope that the two modalities learn in an interactive manner and generate better pseudo-labels. A straightforward way to combine the pseudo-label of the two modalities is to take an average, but we hypothesize that the higher the quality pseudo-labels are, the higher the weights should be assigned.

To verify the effectiveness of the proposed intra-modal consistency weighting, we compare pseudo-label generation based on intra-modal consistency weighting, two-branch pseudo-label averaging, and two-branch logits averaging to generate the final synthetic pseudo-labels. We perform ablation experiments on the four splits of *SET1* and report our results in Table 4. From the results, it can be observed that although averaging over pseudo-labels and averaging over logit to generate pseudo-labels also perform well, our method is superior and outperforms them by a margin. This illustrates the importance of intra-modal consistency weighting for pseudo-label generation.

Qualitative Analysis

In order to better illustrate the effectiveness of our approach, we conducted a quantitative analysis. Specifically, following the approach used by UNO(Fini et al. 2021), IIC(Li et al. 2022a), we used t-SNE(Van der Maaten and Hinton 2008) to visualize the concatenated logits from two classification heads of image branch. Visualization results are shown in Fig.3. It can be seen that our method exhibits better clustering performance on the logits, compared to UNO and IIC. Although mixing is still significant, relatively distinct clusters have emerged.

Discussion

Why not feature concatenation? We assume that the text is not available at test time, so feature concatenation does not apply in our task setup.

Method	SET1				SET2			
	I	II	III	IV	I	II	III	IV
K-means(McQueen 1967)	38.9	37.1	39.2	39.8	30.0	29.8	32.8	33.6
AutoNovel(Han et al. 2021)	38.7	36.8	38.3	39.4	31.3	32.0	37.6	37.0
NCL(Zhong et al. 2021a)	39.6	37.1	39.3	40.8	35.2	34.5	36.8	37.4
UNO(Fini et al. 2021)	43.6	37.5	38.9	46.2	35.8	35.1	36.1	35.6
UNO_MMKD	42.4	37.3	41.5	43.7	34.8	36.6	38.0	36.3
IIC(Li et al. 2022a)	43.8	40.1	44.5	45.6	35.7	35.3	39.2	38.6
ComEx(Yang et al. 2022)	42.7	41.4	40.3	41.1	34.8	35.7	40.8	40.4
Ours	**56.9**	53.7	55.7	**52.6**	50.1	48.8	50.4	47.8
Ours+BYOP(Yang et al. 2023)	56.1	**58.8**	**56.5**	52.4	**50.2**	**49.8**	**51.3**	**50.0**

Table 2: Comparison of state-of-the-art methods on eight splits of SET1 and SET2 using task-aware protocol. Cluster accuracy is reported on the unlabeled test set. The optimal results from the 5 runs are reported in the table. Noting that *UNO_MMKD* means taking the frozen text branch as *teacher* and the image branch as *student* and distilling the knowledge at the logits end.

Method	I			II			III			IV		
	Lab	Unlab	All	Lab	Unlab	All	Lab	Unlab	All	Lab	Unlab	All
UNO(Fini et al. 2021)	38.0	37.5	37.8	39.1	36.6	38.6	40.6	38.9	40.0	43.3	40.2	42.6
IIC(Li et al. 2022a)	39.5	37.6	38.7	40.8	38.5	40.4	41.0	**39.7**	40.6	42.8	37.3	41.6
ComEx(Yang et al. 2022)	41.0	**38.9**	40.1	41.8	38.1	41.1	39.7	36.5	38.6	43.1	39.0	42.2
Ours	**46.3**	38.7	**43.1**	**50.9**	**42.4**	**49.2**	**43.6**	38.3	**41.8**	**55.9**	**44.2**	**53.3**

Table 3: Comparison with some state-of-the-art methods on four splits of SET1 under the task-agnostic protocol. Both the classification accuracy of the labeled test sets("Lab") and the clustering accuracy of the unlabeled test sets("Unlab") are reported.

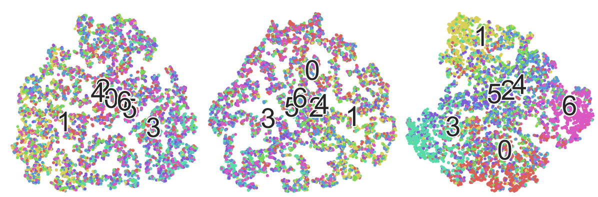
Method	SET1_I	SET1_II	SET1_III	SET1_IV
PL avg.	54.8	53.6	54.8	51.2
Logits avg.	52.7	52.4	51.5	51.8
Weighting	**56.9**	**53.7**	**55.7**	**52.6**

Table 4: Ablation study performed on four splits of SET1 on synthetic pseudo-label generation. *PL avg.* means averaging pseudo-labels from two branches. *Logits avg.* means averaging unlabeled logits from two branches to generate pseudo-labels. *Weighting* means weighting by intra-modal consistency. The results are reported on the unlabeled test set using the task-aware protocol.

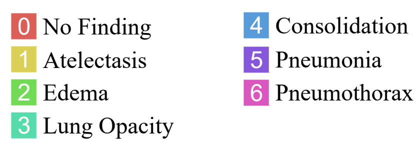
Why not swapping predictions like UNO(Fini et al. 2021)? While text can be viewed as a form of augmentation, we believe it is fundamentally different from augmented images. There is an information difference between image modality and text modality, so we want to achieve both intra-modal and inter-modal optimization when preserving the supervised information from both modalities.

Conclusion

In this paper, we propose a method for NCD in CXR images based on paired images and text. During pretraining, we perform multi-modal contrastive learning on the training set to mitigate the bias towards labeled classes. In the discovery phase, we generate pseudo-labels on both image branch and text branch, respectively, and weight the pseudo-labels by intra-modal consistency. In this way, pseudo-labels that combine information from both modalities are used for training on both branches. Through extensive experiments and analysis, we illustrate the effectiveness of our approach. Our method achieves the best performance on the novel class discovery on CXR images, which is simple but valid.



(a) UNO (b) IIC (c) Ours



(d) Index of numbers, colors to classes.

Figure 3: t-SNE visualization for all classes in SET1.

Acknowledgements

This work was supported by the grants from the National Natural Science Foundation of China (62201014), Beijing Advanced Discipline Construction Project (BMU2019GJJXK001) and PKU-OPPO Innovation Fund (BO202103).

References

Alsentzer, E.; Murphy, J. R.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; and McDermott, M. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Berthelot, D.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Sohn, K.; Zhang, H.; and Raffel, C. 2019a. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019b. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32.

Chen, H.; Tao, R.; Fan, Y.; Wang, Y.; Wang, J.; Schiele, B.; Xie, X.; Raj, B.; and Savvides, M. 2023. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*.

Chi, H.; Liu, F.; Han, B.; Yang, W.; Lan, L.; Liu, T.; Niu, G.; Zhou, M.; and Sugiyama, M. 2021. Meta discovery: Learning to discover novel classes given very limited data. *arXiv preprint arXiv:2102.04002*.

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

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

Han, K.; Rebuffi, S.-A.; Ehrhardt, S.; Vedaldi, A.; and Zisserman, A. 2020. Automatically Discovering and Learning New Visual Categories with Ranking Statistics. In *International Conference on Learning Representations (ICLR)*.

Han, K.; Rebuffi, S.-A.; Ehrhardt, S.; Vedaldi, A.; and Zisserman, A. 2021. AutoNovel: Automatically Discovering and Learning Novel Visual Categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Han, K.; Vedaldi, A.; and Zisserman, A. 2019. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8401–8409.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hsu, Y.-C.; Lv, Z.; and Kira, Z. 2017. Learning to cluster in order to transfer across domains and tasks. *arXiv preprint arXiv:1711.10125*.

Hsu, Y.-C.; Lv, Z.; Schlosser, J.; Odom, P.; and Kira, Z. 2019. Multi-class classification without multi-class labels. *arXiv preprint arXiv:1901.00544*.

Huang, S.-C.; Shen, L.; Lungren, M. P.; and Yeung, S. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3942–3951.

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019a. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317.

Johnson, A. E.; Pollard, T. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Peng, Y.; Lu, Z.; Mark, R. G.; Berkowitz, S. J.; and Horng, S. 2019b. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Kuhn, H. W. 2005. The Hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52(1): 7–21.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896. Atlanta.

Li, W.; Fan, Z.; Huo, J.; and Gao, Y. 2022a. Modeling Inter-Class and Intra-Class Constraints in Novel Class Discovery. *arXiv preprint arXiv:2210.03591*.

Li, Z.; Otholt, J.; Dai, B.; Meinel, C.; Yang, H.; et al. 2022b. A closer look at novel class discovery from the labeled set. *arXiv preprint arXiv:2209.09120*.

Liang, V. W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. Y. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625.

McQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967*, 281–297.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

Sun, Y.; Shi, Z.; Liang, Y.; and Li, Y. 2023. When and How Does Known Class Help Discover Unknown Ones? Provable Understanding Through Spectral Analysis.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wu, J. T.; Agu, N. N.; Lourentzou, I.; Sharma, A.; Paguio, J. A.; Yao, J. S.; Dee, E. C.; Mitchell, W.; Kashyap, S.; Giovannini, A.; et al. 2021. Chest ImaGenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, 478–487. PMLR.

Yang, M.; Wang, L.; Deng, C.; and Zhang, H. 2023. Bootstrap Your Own Prior: Towards Distribution-Agnostic Novel Class Discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3459–3468.

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

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhao, B.; and Han, K. 2021. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *Advances in Neural Information Processing Systems*, 34: 22982–22994.

Zhong, Z.; Fini, E.; Roy, S.; Luo, Z.; Ricci, E.; and Sebe, N. 2021a. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10867–10875.

Zhong, Z.; Zhu, L.; Luo, Z.; Li, S.; Yang, Y.; and Sebe, N. 2021b. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9462–9470.