Towards Realistic Semi-supervised Medical Image Classification

Wenxue Li[1,2*], Lie Ju[1*], Feilong Tang[1*], Peng Xia[3], Xinyu Xiong[4], Ming Hu[1], Lei Zhu[2†], Zongyuan Ge[1†]

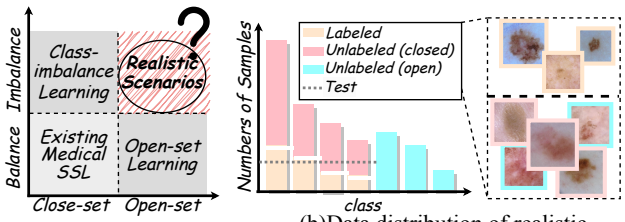
[1] Monash University
[2] The Hong Kong University of Science and Technology (Guangzhou)
[3] UNC-Chapel Hill
[4] Sun Yat-sen University
wxli408@gmail.com, {Lie.Ju1, feilong.tang, zongyuan.ge}@monash.edu, leizhu@ust.hk

Abstract

Existing semi-supervised learning (SSL) approaches follow the idealized closed-world assumption, neglecting the challenges present in realistic medical scenarios, such as open-set distribution and imbalanced class distribution. Although some methods in natural domains attempt to address the open-set problem, they are insufficient for medical domains, where intertwined challenges like class imbalance and small inter-class lesion discrepancies persist. Thus, this paper presents a novel self-recalibrated semantic training framework, which is tailored for SSL in medical imaging by ingeniously harvesting realistic unlabeled samples. Inspired by the observation that certain open-set samples share some similar disease-related representations with in-distribution samples, we first propose an informative sample selection strategy that identifies high-value samples to serve as augmentations, thereby effectively enriching the semantics of known categories. Furthermore, we adopt a compact semantic clustering strategy to address the semantic confusion raised by the above newly introduced open-set semantics. Moreover, to mitigate the interference of class imbalance in open-set SSL, we introduce a less biased dual-balanced classifier with similarity pseudo-label regularization and category-customized regularization. Extensive experiments on a variety of medical image datasets demonstrate the superior performance of our proposed method over state-of-the-art Closed-set and Open-set SSL methods.

Introduction

Semi-supervised Learning (SSL) (Tarvainen and Valpola 2017; Liu et al. 2020; Wang et al. 2022b) offers a promising solution to mitigate the scarcity of medical expert annotations. However, their effectiveness often relies on the assumption of a closed-world scenario. In realistic medical scenarios, open-set distribution (Oliver et al. 2018; Guo et al. 2020; Chen et al. 2020) and imbalanced class distribution (Galdran, Carneiro, and González Ballester 2021; Ju et al. 2022) significantly confuse the model training process, degrades the model performance and further restricts the practical application of SSL. Specifically, on the one hand, unlabeled data often lacks careful selection, as involving experts in meticulous scrutiny of unlabeled data would incur significant costs and undermine the cost-effectiveness of SSL. Thus, the unlabeled images often contain samples from open-set categories. On the other hand, medical datasets naturally encounter the class imbalance issue, primarily caused by the uneven distribution of diseases or lesions (Ju et al. 2022). The performance on minority classes is compromised due to the imbalanced class distribution.



(a) Problem illustration (b)Data distribution of realistic medical scenarios

Figure 1: Illustration of realistic semi-supervised learning for medical scenarios. (a) Problem illustration depicting the combination of open-set and imbalanced class distributions in realistic scenarios. (b) Data distribution in realistic medical scenarios, highlighting the imbalance and diversity in labeled and unlabeled data, with the unlabeled set including both known and unknown categories.

Tremendous efforts have been devoted to addressing the challenges of open-set distribution and class imbalance. For open-set distribution, existing open-set SSL studies (Yu et al. 2020; Saito, Kim, and Saenko 2021; Huang et al. 2021a; Li et al. 2023b) primarily identify and remove open-set samples from unlabeled data to prevent interference from open-set categories during the training process. For the class imbalance challenge, various methods have been proposed to achieve balanced representation learning, including re-sampling (Buda, Maki, and Mazurowski 2018; Dong, Gong, and Zhu 2017; Soltanzadeh, Feizi-Derakhshi, and Hashemzadeh 2023), re-weighting (Lin et al. 2017; Cui et al. 2019; Chen et al. 2023) and self-supervised training (Yang and Xu 2020; Wei et al. 2021; Elbatel et al. 2023). However, these methods do not meet the requirements of realistic medical domain recognition scenarios: a) They tackle these

*These authors contributed equally.
†Corresponding author.

challenges separately, without considering the simultaneous presence of both issues in medical scenarios. b) They often struggle to address the complex interplay between open-set samples and class imbalance.

This paper considers a realistic but more challenging setting: the labeled data consists of predefined known categories, while the unlabeled dataset encompasses both known and previously unknown categories. Additionally, the data presents an imbalanced distribution (Figure 1), forming a semi-supervised training set, and the model is expected to classify all known in a balanced test set. This setting presents two major challenges that existing methods cannot address. **(I) Open-set imbalance problem:** The imbalanced distribution exacerbates the open-set challenge (Sapkota and Yu 2023), as the limited representation of minority classes significantly hinders the model's ability to distinguish between closed-set and open-set samples (Figure 1 (b)). In training progress, the open-set samples cause the model to constrain embeddings of same-class samples with different set combinations towards the class center vector direction. This impedes the utilization of specific information from different sets, resulting in inadequate representation learning. **(II) Intricate inter-class variations problem:** The intricate inter-class variations in medical images, combined with the intertwinement of open-set and class imbalance issues, make it difficult for existing open-set SSL methods to effectively distinguish between open-set and closed-set samples, weakening the model's ability to recognize minority classes.

To alleviate the above challenges, we propose a novel semantic recalibration training framework, which is specifically designed for realistic SSL in the medical domain. In particular, we introduce an informative sample selection strategy to select the informative samples utilized in training, regardless of whether they are closed-set or open-set samples. The selected open-set samples serve as powerful augmentations for enriching the known category semantics. Subsequently, to mitigate the semantic dispersion and confusion caused by newly introduced open-set semantic information during the training, we employ a compact semantic clustering. This approach encourages compact semantic clusters for each category in the feature space, thereby recalibrating the closed-set and open-set semantics and reducing any potential ambiguity. Furthermore, to alleviate the impact of class imbalance in open-set SSL, we introduce a dual-balanced classifier. This classifier comprises two components: a similarity pseudo-label regularization to generate unbiased representation-level pseudo-labels and a category-customized regularization to correct the logit-based pseudo-labels.

To summarize, our contributions are outlined as follows:

- We establish a realistic medical SSL task and we establish a benchmark for evaluating the SSL in realistic medical scenarios. We also present a novel framework to address the challenges of the realistic semi-supervised learning scenario for medical image classification.
- We propose an informative sampling strategy to augment the known semantics using informative open-set data. Moreover, a compact semantic clustering strategy is proposed to achieve semantic recalibration.
- We present a dual-balanced classifier, integrating both similarity pseudo-label regularization and category-customized regularization to mitigate the challenges posed by class imbalance.
- We conduct extensive experiments and the experimental results showcase the state-of-the-art (SOTA) performance and effectiveness of our proposed method.

Related Work

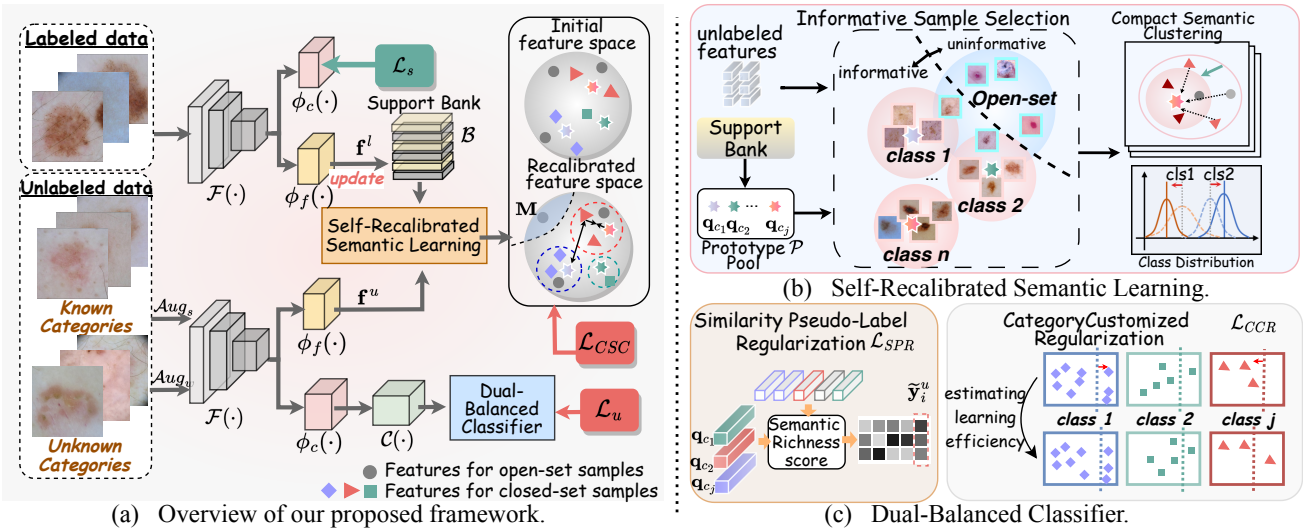
Semi-supervised Learning

Recently, numerous methods have been proposed to tackle the scarcity of annotations (Tang et al. 2024; Zhao et al. 2024), especially in the medical field (Hu et al. 2024; Li et al. 2024b; Wang et al. 2022a; Xia et al. 2024; Song et al. 2024). SSL includes pseudo-labels (Lee et al. 2013; Xie et al. 2020; Zhai et al. 2019) and consistency regularization (Rasmus et al. 2015; Laine and Aila 2017; Tarvainen and Valpola 2017; Ke et al. 2019; Li et al. 2023a). Pseudo-labels-based methods assign labels to unlabeled data based on their confidence scores compared to a predefined threshold value. Consistency regularization methods aim to enforce consistency among different outputs that are generated from the same input images but subjected to various perturbations. This is achieved through techniques such as ensembled model (Rasmus et al. 2015), teacher-student framework (Tarvainen and Valpola 2017), and virtual adversarial training (Miyato et al. 2018), *etc.* Some approaches have combined the strengths of both pseudo-labels and consistency regularization, such as (Sohn et al. 2020; Li, Socher, and Hoi 2020; Zhang et al. 2021; Huang et al. 2023; Li et al. 2024a), resulting in improved performance in SSL.

For the medical domain, it is worth noting that existing SSL methods (Liu et al. 2020, 2021, 2022; Wang et al. 2022b; Gyawali et al. 2020; Ju et al. 2021; Zhou et al. 2019; Wang et al. 2021; Yang et al. 2022b) are all based on the assumption of an ideal closed-set scenario, without considering the challenges that may arise in real-world scenarios. Therefore, it is crucial to develop a SSL method for practical medical scenarios. In this work, we develop a novel SSL method that addresses open-set categories and class imbalance in practical medical scenarios.

Open-set Semi-supervised Learning

The presence of unknown categories in unlabeled data poses significant challenges for traditional SSL algorithms (Oliver et al. 2018; Chen et al. 2020). The issue is commonly referred to as open-set SSL. Unfortunately, the aforementioned standard SSL methods tend to overlook this critical issue. To address the open-set SSL challenge, some methods (Yu et al. 2020; Huang et al. 2021a; Saito, Kim, and Saenko 2021; Li et al. 2023b) have been proposed. These methods aim to recognize and mitigate the negative impacts of open-set samples in unlabeled data to ensure that the classification model is trained with the closed-set samples. For instance, they employ the one-vs-all (OVA) classifier (Saito, Kim, and Saenko 2021), or the feature matching-based classifiers (Huang et al. 2021a) to identify open-set samples.



(a) Overview of our proposed framework. (b) Self-Recalibrated Semantic Learning. (c) Dual-Balanced Classifier.

Figure 2: Overview of our proposed framework, containing a feature extractor $\mathcal{F}(\cdot)$, two feature projection heads $\phi_c(\cdot)$ and $\phi_f(\cdot)$, and a classifier $\mathcal{C}(\cdot)$. The informative sampling strategy and the compact semantic clustering are designed for the open-set challenges. The similarity pseudo-label recalibration and the category customized recalibration are utilized to alleviate the class imbalance issue.

DS3L (Guo et al. 2020) and CAFA (Huang et al. 2021b) assign less weight to open-set samples during training. However, the small inter-class variations and class distribution imbalance (Galdran, Carneiro, and González Ballester 2021; Zhang et al. 2021) in the medical domain significantly exacerbate the difficulty of detecting open-set samples, to the extent that the above open-set SSL methods fail. Therefore, we propose a novel approach to effectively tackle the realistic challenges specific to the medical domain.

Methodology

Problem Definition

Given a labeled set $\mathcal{D}_l = \{(\mathbf{x}_1^l, \mathbf{y}_1^l), \cdots, (\mathbf{x}_n^l, \mathbf{y}_n^l)\}$ and an unlabeled set $\mathcal{D}_u = \{\mathbf{x}_1^u, \cdots, \mathbf{x}_m^u\}$, where $n \ll m$. The class sets of labeled data and unlabeled data are denoted as $\mathcal{C}_l = \{c_1, \cdots, c_k\}$ and $\mathcal{C}_u = \{c_1, \cdots, c_k, \cdots, c_{k+t}\}$. The traditional SSL assumes that labeled and unlabeled data share identical category distribution, *i.e.,* $t = 0$. In a realistic setting, $t > 0$ indicates the presence of additional unknown categories in the unlabeled data. Meanwhile, both $\mathcal{D}_l$ and $\mathcal{D}_u$ present an imbalanced distribution.

Framework Overview

As shown in Figure 2 (a), our proposed framework contains a feature extractor $\mathcal{F}(\cdot)$, two feature projection heads $\phi_c(\cdot)$ and $\phi_f(\cdot)$, and a classifier $\mathcal{C}(\cdot)$. Initially, we introduce a self-recalibrated semantic learning, including an informative sample selection and a compact semantic clustering, to handle the intricate inter-class variations problem in the open-set challenge. Then, to address the open-set imbalance issue, we propose a dual-balanced classifier, which consists of the similarity pseudo-label regularization and the category customized regularization. The overall training objective is

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_u + \alpha \mathcal{L}_{CSC}, \tag{1}$$

where $\mathcal{L}_s$ is the supervised loss on labeled data, $\mathcal{L}_u$ is the unsupervised loss on unlabeled data and $\mathcal{L}_{CSC}$ is our proposed compact semantic clustering regularization. α is the hyper-parameter. The output of classifier is as $\hat{\mathbf{p}}_i = \mathcal{C}(\phi_c(\mathcal{F}(\mathbf{x}_i)))$. For labeled data, the supervised training loss $\mathcal{L}_s$ is

$$\mathcal{L}_s = \frac{1}{B} \sum_{i=1}^{B} \mathbb{H}(\mathbf{y}_i^l, \hat{\mathbf{p}}_i^l), \tag{2}$$

where B is the batch size of labeled data, $\mathbb{H}(\cdot, \cdot)$ is cross-entropy. For unlabeled data, we propose the following strategies to address the issues in the realistic SSL setting.

Self-Recalibrated Semantic Learning

Informative Sample Selection. Previous open-set SSL methods (Yu et al. 2020; Huang et al. 2021a; Fan et al. 2023) take a straightforward approach to addressing the open-set challenge, proposing various strategies to enhance the identification of open-set samples. However, unique challenges in the medical domain make the accurate identification of open-set samples particularly difficult. This has motivated us to design a strategy that can effectively extract valuable features from open-set samples, even without explicitly recognizing them. Unlike in natural domains, we observe that certain open-set medical samples share similar disease-related representations with in-domain samples within the semantic space. For example, in ophthalmology, age-related macular degeneration (AMD) and diabetic retinopathy (DR) share similar and indiscernible lesion features, such as hemorrhage. When AMD is the open-set category in the DR grading task, the semantic information of shared features can be used for augmentation, thereby enhancing robustness. Thus, we explore a new perspective: *"Instead of completely leveraging or discarding them, can we harness beneficial open-set samples during training to achieve favorable*

outcomes?" These open-set samples can serve as augmentations for enriching the known category semantics, whereas previous methods have not discovered this aspect.

Building upon this perspective, we encourage certain hard-to-discriminate open-set samples, denoted as **informative samples**, to be efficient augmentations during training, where open-set semantics can serve as a supplement to known class semantics. Formally, we select the *informative samples* in unlabeled samples during training, using the projection head $\phi_f(\cdot)$ to generate distribution-agnostic feature representations. We utilize a support bank $\mathcal{B}$ to store the features of the known categories and update the prototype pool $\mathcal{P}$ to represent the most representative features of each known category. To be specific, we utilize a support bank $\mathcal{B}$ that stores the features of labeled data as

$$\mathcal{B} = \{\mathbf{f}_i^l \in \mathbb{R}^D | \mathbf{f}_i^l = \mathrm{Norm}(\phi_f(\mathcal{F}(\mathbf{x}_i^l)))\}_{x_i^l \in \mathcal{D}_l}, \tag{3}$$

where D represents the dimension of the projected features. Then we update the prototype pool $\mathcal{P} = \{\mathbf{q}_{c_j}\}_{c_j \in \mathcal{C}_l}$ using normalized features of labeled data in a moving-average fashion, as

$$\mathbf{q}_{c_j}^t = \zeta \mathbf{q}_{c_j}^{t-1} + (1-\zeta)\frac{1}{N_{c_j}} \sum_{i=1}^{N_{c_j}} \mathrm{Norm}(\mathbf{f}_i^l), \tag{4}$$

where $\mathbf{q}_{c_j}^t$ represents the prototype for $j\text{-}th$ category at time step t, N_{c_j} denotes the number of samples in $j\text{-}th$ category. ζ is the momentum factor and $\mathrm{Norm}(\cdot)$ is the normalization operation. By incorporating the support bank and performing momentum-based updates, we could effectively capture the feature representations of the known categories.

Next, we define the semantic richness score of the $i\text{-}th$ sample to select the informative samples, as

$$\mathbf{S}_i = \{\mathbf{s}_j | \mathbf{s}_j = < \mathbf{f}_i^u, \mathbf{q}_{c_j} >, \forall \mathbf{q}_{c_j} \in \mathcal{P}\}, \tag{5}$$

where $\mathbf{f}_i^u = \mathrm{Norm}(\phi_f(\mathcal{F}(\mathbf{x}_i^u)))$, and $< \mathbf{f}_i^u, \mathbf{q}_{c_j} > = \frac{\mathbf{f}_i^{u\top} \cdot \mathbf{q}_{c_j}}{\|\mathbf{f}_i^u\| \|\mathbf{q}_{c_j}\|}$ denotes the cosine similarity. We employ an informative rating mask $\mathbf{M} = \{\mathbf{m}_i^u\}_{i=1}^m$ to ensure the inclusion of samples with significant information during training, where each sample is assigned a signal indicating that if it is informative using the respective $\mathbf{S}_i$, as

$$\mathbf{m}_i = \mathbb{1}((\max \mathbf{S}_i - \min \mathbf{S}_i) \geq \sigma), \tag{6}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function and σ is the hyper-parameter threshold. The selected samples exhibit high similarity to a known category while demonstrating lower similarity to another category, ensuring that they have rich semantic information. The informative samples may contain open-set samples that share characteristics with a specific known category. Despite being outside the known class distribution, the open-set semantics can provide valuable information and serve as augmentations to the specific categories. Incorporating such samples narrows the gap between the target closed-set distribution and the complex real-world distribution, improving the model's ability to handle variations and unforeseen inputs.

Compact Semantic Clustering. The introduced informative open-set augmentations may cause semantic dispersion and confusion during training. To alleviate the semantic dispersion, we employ a compact semantic clustering regularization to recalibrate semantics, as

$$\mathcal{L}_{CSC} = -\frac{1}{\lambda B} \sum_{i=1}^{\lambda B} \mathbf{m}_i \cdot \log \frac{\exp(\frac{\mathbf{f}_i^u \cdot \mathbf{p}^+}{\tau})}{\exp(\frac{\mathbf{f}_i^u \cdot \mathbf{p}^+}{\tau}) + \sum_{\mathbf{p}^- \in \mathcal{P}} \exp(\frac{\mathbf{f}_i^u \cdot \mathbf{p}^-}{\tau})}, \tag{7}$$

where λ is the hyper-parameter to control the batch size of unlabeled data. $\mathbf{p}^+$ is the most similar prototype vector to the sample and $\mathbf{p}^-$ refers to the other prototypes in $\mathcal{P}$. This design maximizes similarity between positive pairs and minimizes the similarity between negative pairs. Positive pairs are informative samples and their closest prototypes, while negative pairs are informative samples and other prototypes (excluding the closest prototypes). This encourages tighter clusters per category in the feature space, resulting in clearer decision boundaries for closed-set samples and bringing open-set semantics closer to their corresponding classes.

Dual-Balanced Classifier

To address the imbalanced class distribution, we develop a dual-balanced classifier (DRC), aiming to regularize the outputs using a logit-based classifier and representation-based matching scores. DRC consists of two main components: similarity pseudo-label regularization (SPR) and category-customized regularization (CCR).

Similarity Pseudo-Label Regularization. Recent studies have demonstrated the potential of representation-level similarity-based classification in dealing with imbalanced distributions (Rebuffi et al. 2017; Li, Xiong, and Hoi 2021; Kang et al. 2020; Oh, Kim, and Kweon 2022). This approach generates labels by measuring the similarity between category centers and feature representations (Snell, Swersky, and Zemel 2017). Outputs obtained through similarity measuring exhibit lower biases in class distribution compared to those generated by the logit-based classifiers (Oh, Kim, and Kweon 2022). Thus, we recalibrate the outputs by leveraging representation-level similarity measurements, which generate labels by measuring the similarity between category centers and feature representations. The recalibrated pseudo-label of the $i\text{-}th$ sample is based on semantic richness score $\mathbf{S}_i$, as $\widetilde{\mathbf{y}}_i^u = \arg\max \mathbf{S}_i$. We assign the informative samples with $\widetilde{\mathbf{y}}_i^u$ and the similarity pseudo-label regularization is formulated as

$$\mathcal{L}_{SPR} = \frac{1}{\lambda B} \sum_{i=1}^{\lambda B} \mathbf{m}_i \cdot \mathbb{H}(\widetilde{\mathbf{y}}_i^u, \hat{\mathbf{p}_i^u}). \tag{8}$$

Category Customized Regularization. Pseudo-labeled based methods (Lee et al. 2013; Sohn et al. 2020) adopt a fixed threshold, ignoring the learning efficiency of individual categories. This can result in a model that disproportionately prioritizes the majority or easier-to-learn classes while neglecting the minority or more challenging ones. Thus, we introduce an attentive category customized regularization

technique. This approach involves setting individual thresholds for each class, allowing for tailored adjustments based on the unique characteristics and difficulty levels of each category. For informative sample, we estimate the leaning efficiency for each category using the confident sample ratio $\mathcal{E}$, as

$$\mathcal{E} = \{e_{c_j} | e_{c_j} = \sum_{i=1}^{N_{c_j}} \mathbb{1}(\max(\hat{\mathbf{p}}_i^u) > \epsilon), \forall c_j \in \mathcal{C}_l\} \quad (9)$$

when $\arg\max(\hat{\mathbf{p}}_i^u) = c_j$, where ϵ is a pre-defined threshold and we empirically adopt 0.95 following (Sohn et al. 2020). The threshold for j-th class is $\epsilon_{c_j} = \frac{e_{c_j}}{\max_{c_j \in \mathcal{C}_l} e_{c_j}} \epsilon$. Next, we employ this adaptable threshold to refine the selection of samples for each category involved in the learning process. The category customized regularization $\mathcal{L}_{CCR}$ is defined as

$$\mathcal{L}_{CCR} = \frac{1}{\lambda B} \sum_{i=1}^{\lambda B} \mathbb{1}(\max(\hat{\mathbf{p}}_i^u) > \epsilon_{c_j}) \cdot \mathbf{m}_i \cdot \mathbb{H}(\hat{\mathbf{y}}_i^u, \hat{\mathbf{p}}_i^u), \quad (10)$$

where $\hat{\mathbf{y}}_i^u = \arg\max(\mathcal{C}(\phi_c(\mathcal{F}(\mathbf{x}_i^u))))$ and $\arg\max(\hat{\mathbf{p}}_i^u) = c_j$. It guarantees that the training data includes a higher proportion of samples from more challenging or minority classes while imposing stricter criteria for the selection from easier or majority classes.

Overall, the unsupervised loss is composed by $\mathcal{L}_{CCR}$ and $\mathcal{L}_{SPR}$, as

$$\mathcal{L}_u = \mathcal{L}_{CCR} + \beta \mathcal{L}_{SPR}. \quad (11)$$

Experiments

Dataset

We validate our proposed method on diverse datasets comprising multiple modalities, including dermatology, ophthalmology, and endoscopy.

- **Dermatology.** We adopt ISIC-2019 (Combalia et al. 2022), which is a comprehensive collection of dermatology images, encompassing 8 kinds of skin lesions. In our experiments, 4 classes (MEL, NV, BCC, and BKL) are considered as known classes and the remaining 4 classes (AK, DF, VASC, SCC) are chosen as unknown classes.
- **Ophthalmology.** We utilize APTOS-2019 (Karthick and Sohier 2019), which consists of retinal fundus images for grading diabetic retinopathy (DR). APTOS-2019 includes 5 severity levels of the disease: normal, mild non-proliferative DR (NPDRI), moderate non-proliferative DR (NPDRII), severe non-proliferative DR (NPDRIII), and proliferative DR (PDR). To create a more challenging and realistic scenario, we incorporate samples from the iAMD-Challenge (Fang et al. 2022) dataset, which focuses on age-related macular degeneration (AMD), into the APTOS-2019. By introducing AMD samples as the unknown class, we aim to simulate a scenario where DR and AMD images can be easily confused due to their shared retinal abnormalities.
- **Endoscopy.** We employ HyperKvasir (Borgli et al. 2020), a dataset comprising comprehensive gastrointestinal endoscopic images encompassing a total of 15 classes (excluding grading subclasses). To create a distinction, we rank the classes in descending order according to the number of samples per class and designate the top 8 classes as known, treating the remaining classes as unknown.

Notably, all datasets we utilized exhibit imbalanced class distribution. We consider labeled ratio $\gamma \in \{5\%, 10\%, 20\%\}$ for ISIC-2019, $\gamma \in \{10\%, 20\%\}$ for APTOS-2019, and $\gamma \in \{1\%, 2\%\}$ for HyperKvasir. To construct training data, we sample γ (%) samples from each known class as the labeled dataset, while the remaining samples are used to form the unlabeled data set. We establish the balanced validation set and test set with known classes for each dataset to ensure fair evaluation of the learning performance for every category.

Experimental Setup

We adopt ResNet-50 (He et al. 2016) as the backbone architecture. All images are resized to 224×224. We train the model for 20,000 iterations. To update prototypes, we train the model with only the supervised training manner for the first 200 iterations, without incorporating any other strategies. We adopt the Adam optimizer with a batch size of 64. The hyper-parameter λ which controls the ratio of unlabeled data in each batch, is set to 3. The learning rate is set to 0.0001 and adjusted using the cosine decay strategy. All the experiments are implemented on two NVIDIA RTX4090 GPUs. To ensure the fairness of our comparative study, we use the same basic hyper-parameters for all the methods. The temperature hyper-parameter τ in Eq. 7 is set to 0.07 empirically (He et al. 2020).

Comparison Study

We compare our proposed method with various methods, including closed-set SSL methods and open-set SSL methods. Specifically, for closed-set SSL methods, we compare the methods including Pi-Model (Rasmus et al. 2015), Mean Teacher (Tarvainen and Valpola 2017), MixMatch (Berthelot et al. 2019), FixMatch (Sohn et al. 2020), FlexMatch (Zhang et al. 2021), and CCSSL (Yang et al. 2022a). For open-set SSL methods, we evaluate OpenMatch (Saito, Kim, and Saenko 2021), MTCF (Yu et al. 2020), and T2T (Huang et al. 2021a). We run the comparison method experiments three times using different random seeds and report the mean standard deviation results across all the datasets.

As shown in Table 1, our proposed methods outperform other approaches and achieve new state-of-the-art performance across all the datasets. Notably, all SSL methods demonstrate significant improvements compared to the supervised training approach, except for MixMatch. In realistic settings, our proposed method outperforms those closed-set SSL methods.

Ablation Study

Contribution of important components. We first compare some SSL methods in Table 2 under closed-set setting (unlabeled data contain open-set samples) and open-set setting (unlabeled data does not contain open-set samples). These methods totally utilize or discard open-set samples

Method	Dermatology (ISIC-2019)			Ophthalmology (APTOS-2019)			Endoscopy (HyperKvasir)	
fully supervised	83.5			65.0			96.6	
γ	5%	10%	20%	5%	10%	20%	1%	2%
supervised	62.83±1.47	67.17±0.78	70.93±0.61	49.33±0.62	52.00±2.16	56.50±1.22	83.97±0.95	88.77±0.68
Pi-Model	63.21±1.73	67.40±0.91	71.37±0.54	49.33±1.18	54.50±2.12	58.17±1.25	83.97±1.58	89.91±1.22
Mean Teacher	64.40±1.15	68.73±1.46	71.47±0.84	51.50±1.08	54.67±2.49	58.50±1.63	84.50±1.10	90.83±1.30
MixMatch	59.70±0.65	66.10±0.54	67.93±0.26	48.67±3.66	54.00±1.78	57.50±0.82	86.71±1.33	89.41±1.27
FixMatch	64.93±0.56	68.60±0.73	72.70±0.36	51.83±1.31	55.83±2.25	59.67±0.94	86.24±0.54	94.31±0.21
FlexMatch	67.10±0.54	70.83±0.12	73.53±0.82	53.67±0.85	57.17±1.70	61.00±0.41	88.21±1.19	93.63±1.34
CCSSL	65.90±0.51	69.23±1.39	72.43±0.62	51.33±1.31	56.83±0.94	59.83±1.03	86.64±0.83	94.03±0.51
OpenMatch	64.30±1.23	67.47±1.30	71.00±1.06	46.83±1.53	50.33±1.26	52.83±1.04	82.36±3.19	88.30±2.33
MTCF	64.17±1.03	68.37±0.49	69.93±0.35	51.00±2.65	54.60±2.08	59.17±2.75	86.76±0.64	93.79±0.75
T2T	66.17±1.08	68.63±1.16	71.87±0.90	53.60±0.36	56.33±1.53	58.00±1.32	86.73±1.70	93.28±0.46
Ours	**69.19**±0.34	**72.27**±0.78	**74.90**±0.70	**56.83**±2.95	**59.00**±1.47	**62.50**±1.22	**89.59**±1.29	**94.41**±0.82

Table 1: Comparison with SOTA SSL methods on different datasets, including dermatology dataset (ISIC-2019 (Combalia et al. 2022)), ophthalmology dataset (APTOS-2019 (Karthick and Sohier 2019)) and endoscopy dataset (HyperKvasir (Borgli et al. 2020)) under different labeled ratio γ.

Setting	Pi	Pseudo	MT	FM
Closed-set	65.8	64.8	66.9	65.2
Open-set	60.8	63.2	62.9	64.7

Table 2: Comparision results on ISIC-2019 ($\gamma = 5\%$) when adopting closed-set setting and open-set setting.

ISS	CSC	DRC	ISIC-2019	APTOS-2019	Hyper-Kvasir
✗	✗	✗	64.7	53.0	87.0
✓	✗	✗	67.8	54.5	87.3
✗	✓	✗	65.3	54.5	87.5
✗	✗	✓	66.4	54.0	87.2
✓	✓	✗	67.9	55.0	87.9
✓	✓	✓	69.2	56.8	89.6

Table 3: Model performance on ISIC-2019 ($\gamma = 5\%$), APTOS-2019 ($\gamma = 5\%$), and HyperKvasir ($\gamma = 1\%$) when using different components in our framework, including informative sample selection (ISS), compact semantic clustering (CSC), dual-balanced classifier (DRC).

Components	MEL	NV	BCC	BKL
w/o SPR + w/o CCR	48.0	86.0	77.2	47.6
w/o SPR + w/ CCR	58.0	78.4	82.0	52.8
w/ SPR + w/o CCR	48.8	87.6	78.0	55.6
w/ SPR + w/ CCR	59.2	80.8	81.2	54.4

Table 4: Model performance on the ISIC-2019 dataset ($\gamma = 5\%$) when using different components in dual-balanced classifier, including similarity pseudo-Label regularization (SPR) and category customized regularization (CCR).

in closed- and open-set settings. We can observe that introducing all open-set samples in the unlabeled data leads to a performance decline. Then, Table 3 presents a comprehensive analysis of the contributions of the components in our approach. The baseline with informative sample selection (ISS), which utilizes open-set semantics as effective augmentations, achieves improvement on all these datasets. By incorporating open-set samples that resemble a specific class, the augmented training data provides additional information to the model, potentially leading to improved classification accuracy for that class. Leveraging open-set samples as a valuable source of data augmentation holds immense potential for enhancing generalization performance. Meanwhile, when only adopting the compact semantic clustering (CSC), all open-set samples are forcibly used to augment the semantics of a specific category and such chaotic augmentations can still bring about marginal improvement. When we combine ISS and CSC, the model achieves better performance. Moreover, we measure the accuracy of each category separately to assess the effectiveness of the components in dual-balanced classifier (shown in Table 4). It can be observed that both SPR and CCR improve the performance of the categories with low learning efficiency.

Comparison with class-imbalance methods. We alternate the DRC with Focal Loss and Class-Balanced Loss (CB Loss) to demonstrate the effectiveness of DRC. The experimental results in Figure 3 indicate that alleviating the class imbalance issue can boost overall model performance in the medical domain. The results also indicate that our DRC outperforms the above methods.

Detailed Analysis

In this section, we conduct a detailed analysis from several perspectives to explore the effectiveness of our method.

The Impact of Hyper-parameters. We investigate the impact of hyper-parameters in $\mathcal{L}$ (Eq. 1) on ISIC-2019. As

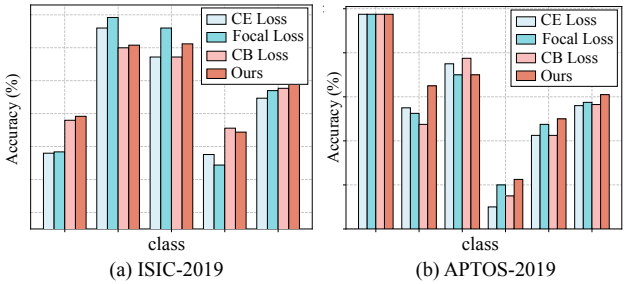


Figure 3: Experimental results of each class using CE Loss, Focal Loss, CB loss and our proposed DRC for ISIC-2019 ($\gamma = 5\%$) and APTOS-2019 ($\gamma = 10\%$).

Loss weight	0.1	0.2	0.5	1.0
α	65.2	67.8	68.1	68.9
β	68.9	68.8	68.7	67.9

Table 5: Model performance (Accuracy (%)) under different loss weight α of $\mathcal{L}_{CSC}$ and β of $\mathcal{L}_{SPR}$ on ISIC-2019 ($\gamma = 5\%$).

shown in Table 5, the performance steadily improves as α increases from 0.1 to 1.0, with optimal performance achieved at $\alpha = 1$. Moreover, the model is not sensitive to β and achieves the best performance when $\beta = 0.1$.

T-SNE Visualization. The t-SNE visualization results of FixMatch (Sohn et al. 2020) and our proposed method are shown in Figure 4. The red dots denote the feature representations of open-set samples while the dots in other colors denote the ID samples. Obviously, for FixMatch, open-set samples distribute more sparsely and cause semantic dispersion within the feature space. Our approach consciously excludes some open-set samples that lack informative semantics and naturally assigns the remaining open-set samples to a preferred category. These open-set semantics augment the ID semantics in training. Moreover, by encouraging compact clustering through $\mathcal{L}_{CSC}$, we facilitate better separations for different classes, leading to improved classification performance.

Interpretability Analysis of Augmentations with Open-set Semantics. In Figure 5, we present the visualization of lesion details and the explanation maps generated by Grad-CAM (Selvaraju et al. 2017) technique. The lesion details provide a reasonable explanation for the augmentations using open-set semantics. Specifically, we can observe small inter-class discrepancies between NPDRII lesions (hard exudates) and AMD lesions (drusen). Drusen, which exhibits a similar appearance to hard exudates, enriches the representations of NPDRII. Furthermore, it is evident that both open-set and closed-set samples exhibit similar lesion types, namely hard exudates and hemorrhage. Therefore, it is understandable why samples from AMD are misclassified as NPDRIII category, and why certain AMD samples are utilized to expand the semantics of NPDRIII. By introducing the informative open-set semantics in the training process, the model can learn to generalize better and potentially become more robust to similar open-set samples.

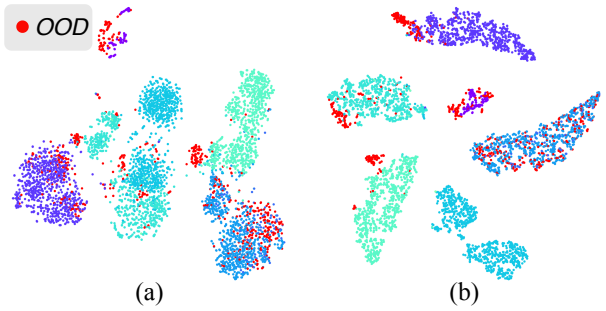


Figure 4: Visualization results of t-SNE of (a) FixMatch (Sohn et al. 2020) and (b) our proposed method for HyperKvasir ($\gamma = 5\%$).

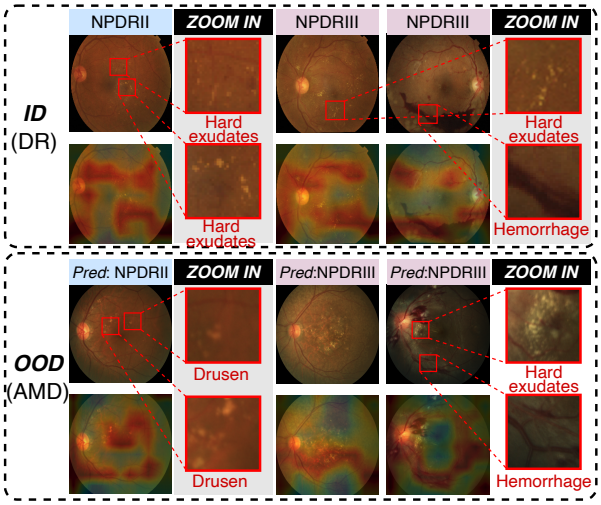


Figure 5: Visualization results of ID (DR) and informative OOD samples (AMD) in the field of ophthalmology illustrate how OOD semantics serve as semantic augmentations for ID categories.

Conclusion

In this paper, we propose a novel semantic recalibration training framework for semi-supervised medical image classification that aims to mitigate the challenges in realistic scenarios. Specifically, we propose an informativeness sampling strategy to leverage useful OOD samples as strong semantic augmentations during training. Additionally, we introduced a compact semantic clustering regularization to achieve semantic recalibration. Moreover, we introduce a dual-balanced classifier, consisting of both the similarity pseudo-label regularization and category-customized regularization to tackle the class-imbalance challenge. We conduct experiments under our established realistic benchmark. The experimental results demonstrate the effectiveness and superior performance of our proposed method in real-world scenarios.

Acknowledgments

The work was supported by Monash-Airdoc Research.

References

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, volume 32.

Borgli, H.; Thambawita, V.; Smedsrud, P. H.; Hicks, S.; Jha, D.; Eskeland, S. L.; Randel, K. R.; Pogorelov, K.; Lux, M.; Nguyen, D. T. D.; et al. 2020. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1): 283.

Buda, M.; Maki, A.; and Mazurowski, M. A. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106: 249–259.

Chen, X.; Zhou, Y.; Wu, D.; Yang, C.; Li, B.; Hu, Q.; and Wang, W. 2023. Area: adaptive reweighting via effective area for long-tailed classification. In *ICCV*, 19277–19287.

Chen, Y.; Zhu, X.; Li, W.; and Gong, S. 2020. Semi-supervised learning under class distribution mismatch. In *AAAI*, volume 34, 3569–3576.

Combalia, M.; Codella, N.; Rotemberg, V.; Carrera, C.; Dusza, S.; Gutman, D.; Helba, B.; Kittler, H.; Kurtansky, N. R.; Liopyris, K.; et al. 2022. Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge. *The Lancet Digital Health*, 4(5): e330–e339.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *CVPR*, 9268–9277.

Dong, Q.; Gong, S.; and Zhu, X. 2017. Class rectification hard mining for imbalanced deep learning. In *ICCV*, 1851–1860.

Elbatel, M.; Wang, H.; Mart, R.; Fu, H.; and Li, X. 2023. Federated model aggregation via self-supervised priors for highly imbalanced medical image classification. In *MICCAI*, 334–346.

Fan, Y.; Kukleva, A.; Dai, D.; and Schiele, B. 2023. SSB: Simple but Strong Baseline for Boosting Performance of Open-Set Semi-Supervised Learning. In *ICCV*, 16068–16078.

Fang, H.; Li, F.; Fu, H.; Sun, X.; Cao, X.; Lin, F.; Son, J.; Kim, S.; Quellec, G.; Matta, S.; et al. 2022. Adam challenge: Detecting age-related macular degeneration from fundus images. *IEEE Transactions on Medical Imaging*, 41(10): 2828–2847.

Galdran, A.; Carneiro, G.; and González Ballester, M. A. 2021. Balanced-mixup for highly imbalanced medical image classification. In *MICCAI*, 323–333.

Guo, L.-Z.; Zhang, Z.-Y.; Jiang, Y.; Li, Y.-F.; and Zhou, Z.-H. 2020. Safe deep semi-supervised learning for unseen-class unlabeled data. In *ICML*, 3897–3906.

Gyawali, P. K.; Ghimire, S.; Bajracharya, P.; Li, Z.; and Wang, L. 2020. Semi-supervised medical image classification with global latent mixing. In *MICCAI*, 604–613.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hu, M.; Xia, P.; Wang, L.; Yan, S.; Tang, F.; Xu, Z.; Luo, Y.; Song, K.; Leitner, J.; Cheng, X.; et al. 2024. Ophnet: A large-scale video benchmark for ophthalmic surgical workflow understanding. In *ECCV*.

Huang, J.; Fang, C.; Chen, W.; Chai, Z.; Wei, X.; Wei, P.; Lin, L.; and Li, G. 2021a. Trash to treasure: Harvesting ood data with cross-modal matching for open-set semi-supervised learning. In *CVPR*, 8310–8319.

Huang, Z.; Shen, L.; Yu, J.; Han, B.; and Liu, T. 2023. Flatmatch: Bridging labeled data and unlabeled data with cross-sharpness for semi-supervised learning. In *NeurIPS*, volume 36, 18474–18494.

Huang, Z.; Xue, C.; Han, B.; Yang, J.; and Gong, C. 2021b. Universal semi-supervised learning. In *NeurIPS*, volume 34, 26714–26725.

Ju, L.; Wang, X.; Zhao, X.; Lu, H.; Mahapatra, D.; Bonnington, P.; and Ge, Z. 2021. Synergic adversarial label learning for grading retinal diseases via knowledge distillation and multi-task learning. *IEEE Journal of Biomedical and Health Informatics*, 25(10): 3709–3720.

Ju, L.; Wu, Y.; Wang, L.; Yu, Z.; Zhao, X.; Wang, X.; Bonnington, P.; and Ge, Z. 2022. Flexible sampling for long-tailed skin lesion classification. In *MICCAI*, 462–471.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2020. Decoupling representation and classifier for long-tailed recognition. In *ICLR*.

Karthick, M.; and Sohier, D. 2019. APTOS 2019 Blindness Detection. https://kaggle.com/competitions/aptos2019-blindness-detection/.

Ke, Z.; Wang, D.; Yan, Q.; Ren, J.; and Lau, R. W. 2019. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *CVPR*, 6728–6736.

Laine, S.; and Aila, T. 2017. Temporal Ensembling for Semi-Supervised Learning. In *ICLR*.

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, J.; Socher, R.; and Hoi, S. C. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*.

Li, J.; Xiong, C.; and Hoi, S. C. 2021. Mopro: Webly supervised learning with momentum prototypes. In *ICLR*.

Li, M.; Wu, R.; Liu, H.; Yu, J.; Yang, X.; Han, B.; and Liu, T. 2024a. Instant: Semi-supervised learning with instance-dependent thresholds. In *NeurIPS*, volume 36.

Li, W.; Lu, W.; Chu, J.; Tian, Q.; and Fan, F. 2023a. Confidence-guided mask learning for semi-supervised medical image segmentation. *Computers in Biology and Medicine*, 165: 107398.

Li, W.; Xiong, X.; Xia, P.; Ju, L.; and Ge, Z. 2024b. TP-DRSeg: improving diabetic retinopathy lesion segmentation with explicit text-prompts assisted SAM. In *MICCAI*, 743–753.

Li, Z.; Qi, L.; Shi, Y.; and Gao, Y. 2023b. Iomatch: Simplifying open-set semi-supervised learning with joint inliers and outliers utilization. In *ICCV*, 15870–15879.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *CVPR*, 2980–2988.

Liu, F.; Tian, Y.; Chen, Y.; Liu, Y.; Belagiannis, V.; and Carneiro, G. 2022. ACPL: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. In *CVPR*, 20697–20706.

Liu, Q.; Yang, H.; Dou, Q.; and Heng, P.-A. 2021. Federated semi-supervised medical image classification via inter-client relation matching. In *MICCAI*, 325–335.

Liu, Q.; Yu, L.; Luo, L.; Dou, Q.; and Heng, P. A. 2020. Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE transactions on medical imaging*, 39(11): 3429–3440.

Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE TPAMI*, 41(8): 1979–1993.

Oh, Y.; Kim, D.-J.; and Kweon, I. S. 2022. Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. In *CVPR*, 9786–9796.

Oliver, A.; Odena, A.; Raffel, C. A.; Cubuk, E. D.; and Goodfellow, I. 2018. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, volume 31.

Rasmus, A.; Berglund, M.; Honkala, M.; Valpola, H.; and Raiko, T. 2015. Semi-supervised learning with ladder networks. In *NeurIPS*, volume 28.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *CVPR*, 2001–2010.

Saito, K.; Kim, D.; and Saenko, K. 2021. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. In *NeurIPS*, volume 34, 25956–25967.

Sapkota, H.; and Yu, Q. 2023. Adaptive Robust Evidential Optimization For Open Set Detection from Imbalanced Data. In *ICLR*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *NeurIPS*, volume 30.

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. In *NeurIPS*, volume 33, 596–608.

Soltanzadeh, P.; Feizi-Derakhshi, M. R.; and Hashemzadeh, M. 2023. Addressing the class-imbalance and class-overlap problems by a metaheuristic-based under-sampling approach. *PR*, 143: 109721.

Song, W.; Zhao, H.; Ding, P.; Cui, C.; Lyu, S.; Fan, Y.; and Wang, D. 2024. GeRM: A Generalist Robotic Model with Mixture-of-experts for Quadruped Robot. In *IROS*.

Tang, F.; Xu, Z.; Qu, Z.; Feng, W.; Jiang, X.; and Ge, Z. 2024. Hunting Attributes: Context Prototype-Aware Learning for Weakly Supervised Semantic Segmentation. In *CVPR*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, volume 30.

Wang, J.; Huang, Q.; Tang, F.; Meng, J.; Su, J.; and Song, S. 2022a. Stepwise feature fusion: Local guides global. In *MICCAI*.

Wang, X.; Chen, H.; Xiang, H.; Lin, H.; Lin, X.; and Heng, P.-A. 2021. Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification. *Medical image analysis*, 70: 102010.

Wang, Z.; Ye, M.; Zhu, X.; Peng, L.; Tian, L.; and Zhu, Y. 2022b. Metateacher: Coordinating multi-model domain adaptation for medical image classification. In *NeurIPS*, volume 35, 20823–20837.

Wei, C.; Sohn, K.; Mellina, C.; Yuille, A.; and Yang, F. 2021. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *CVPR*, 10857–10866.

Xia, P.; Hu, M.; Tang, F.; Li, W.; Zheng, W.; Ju, L.; Duan, P.; Yao, H.; and Ge, Z. 2024. Generalizing to unseen domains in diabetic retinopathy with disentangled representations. In *MICCAI*, 427–437.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-training with noisy student improves imagenet classification. In *CVPR*, 10687–10698.

Yang, F.; Wu, K.; Zhang, S.; Jiang, G.; Liu, Y.; Zheng, F.; Zhang, W.; Wang, C.; and Zeng, L. 2022a. Class-aware contrastive semi-supervised learning. In *CVPR*, 14421–14430.

Yang, Q.; Liu, X.; Chen, Z.; Ibragimov, B.; and Yuan, Y. 2022b. Semi-supervised Medical Image Classification with Temporal Knowledge-Aware Regularization. In *MICCAI*, 119–129.

Yang, Y.; and Xu, Z. 2020. Rethinking the value of labels for improving class-imbalanced learning. In *NeurIPS*, volume 33, 19290–19301.

Yu, Q.; Ikami, D.; Irie, G.; and Aizawa, K. 2020. Multi-task curriculum framework for open-set semi-supervised learning. In *ECCV*, 438–454.

Zhai, X.; Oliver, A.; Kolesnikov, A.; and Beyer, L. 2019. S4l: Self-supervised semi-supervised learning. In *CVPR*, 1476–1485.

Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; and Shinozaki, T. 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *NeurIPS*, volume 34, 18408–18419.

Zhao, X.; Tang, F.; Wang, X.; and Xiao, J. 2024. Sfc: Shared feature calibration in weakly supervised semantic segmentation. In *AAAI*.

Zhou, Y.; He, X.; Huang, L.; Liu, L.; Zhu, F.; Cui, S.; and Shao, L. 2019. Collaborative learning of semi-supervised segmentation and classification for medical images. In *CVPR*, 2079–2088.