PEFAT: Boosting Semi-supervised Medical Image Classification via Pseudo-loss Estimation and Feature Adversarial Training

Qingjie Zeng[1*] Yutong Xie[2*] Zilin Lu[1] Yong Xia[1†]
[1] School of Computer Science and Engineering, Northwestern Polytechnical University, China
[2] The University of Adelaide, Australia
maxwell@mail.nwpu.edu.cn, yutong.xie678@gmail.com, luzl@mail.nwpu.edu.cn, yxia@nwpu.edu.cn

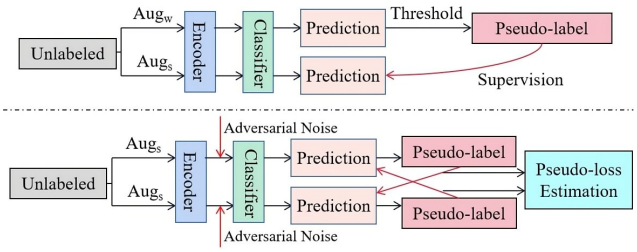
Abstract

*Pseudo-labeling approaches have been proven beneficial for semi-supervised learning (SSL) schemes in computer vision and medical imaging. Most works are dedicated to finding samples with high-confidence pseudo-labels from the perspective of model predicted probability. Whereas this way may lead to the inclusion of incorrectly pseudo-labeled data if the threshold is not carefully adjusted. In addition, low-confidence probability samples are frequently disregarded and not employed to their full potential. In this paper, we propose a novel **P**seudo-loss **E**stimation and **F**eature **A**dversarial **T**raining semi-supervised framework, termed as PEFAT, to boost the performance of multi-class and multi-label medical image classification from the point of loss distribution modeling and adversarial training. Specifically, we develop a trustworthy data selection scheme to split a high-quality pseudo-labeled set, inspired by the dividable pseudo-loss assumption that clean data tend to show lower loss while noise data is the opposite. Instead of directly discarding these samples with low-quality pseudo-labels, we present a novel regularization approach to learn discriminate information from them via injecting adversarial noises at the feature-level to smooth the decision boundary. Experimental results on three medical and two natural image benchmarks validate that our PEFAT can achieve a promising performance and surpass other state-of-the-art methods. The code is available at https://github.com/maxwell0027/PEFAT.*

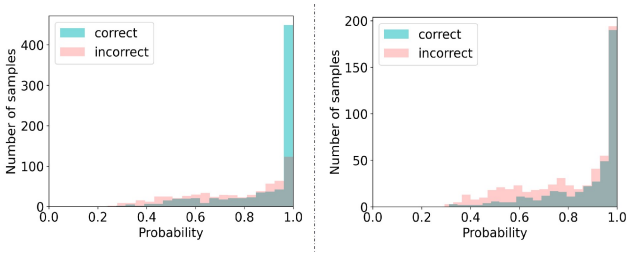
1. Introduction

Deep learning has achieved remarkable success in various computer vision tasks [4–6, 13, 15, 19]. This success has also made practical applications more accessible, including medical image analysis (MIA) [10, 21, 30, 31, 35, 39]. However, unlike computer vision, annotating a large-scale medical image dataset requires expert knowledge and is time-consuming and costly. Alternatively, unlabeled data can be collected from clinical sites in a more available way, thereby mitigating the cost of data annotation by leveraging these unlabeled data.

Semi-supervised Learning (SSL) has drawn a lot of attention due to its superior performance, by only leveraging limited labeled data and a vast number of unlabeled data. Under the SSL setting, it is critical to mine adequate infor-



(a) Illustration of traditional SSL methods (top) and our method (bottom).



(b) Probability distribution of labeled data (left) and validation data (right), when using the warm-upped model on ISIC2018 dataset.

Figure 1. (a) shows the main difference between our method and other SSL methods, our method selects high-quality pseudo-labeled data by pseudo-loss estimation, and also injects feature-level adversarial noises for better unlabeled data mining; (b) indicates the phenomenon that clean pseudo-labeled set is hard to collect when using probability-based threshold, which is mainly attributed to the over-confident prediction.

*Equal contribution. †Y. Xia is the corresponding author. This work was supported in part by the National Natural Science Foundation of China under Grants 62171377, and in part by the Key Research and Development Program of Shaanxi Province, China, under Grant 2022GY-084.

mation from unlabeled data. In the existing SSL methods, pseudo-labeling [14, 22, 29, 37, 38] and consistency regularization [12, 24, 25] are the mainstream. The former focuses on finding confident pseudo-labels for re-training, and the latter aims to improve the robustness of the model by keeping one logical distribution similar to the other.

However, most SSL methods encounter two issues. First, unreliable pseudo-labels are a problem with the threshold-selecting data method based on predicted probability as it often introduces numerous incorrect pseudo-labels due to confirmation bias. As illustrated in Figure 1b, unlabeled data with both correct and incorrect pseudo-labels follow similar probability distributions. Second, informative unlabeled samples are underutilized as unselected data with low probabilities typically cluster around the decision boundary. Recent studies [1, 17] have found that neural networks tend to fit clean data first and then memorize noise data during training, resulting in lower loss for clean data and higher loss for noise data in early stages of training. Furthermore, some works [9, 25] have investigated the effects of adversarial training under semi-supervised settings, which show potential for learning from low-quality pseudo-labeled data. All these findings pave the way towards solving the aforementioned problems.

In this paper, we propose a novel SSL method called **P**seudo-loss **E**stimation and **F**eature **A**dversarial **T**raining (PEFAT) for multi-class and multi-label medical image classification. First, we introduce a new estimation scheme for reliable pseudo-labeled data selection from the perspective of pseudo-loss distribution. It is motivated by our argument that there is a dividable loss distribution between correct and incorrect pseudo-labeled data. Specifically, we first warm up the model on training data with contrastive learning, in order to learn unbiased representation. Then we set up a two-component Gaussian Mixture Model (GMM) [28] to learn prior loss distribution on labeled data. In the procedure of pseudo-labeled data selection, we feed cross pseudo-loss to the fitted GMM and obtain the trustworthy pseudo-labeled data with posterior probability. Second, we propose a feature adversarial training (FAT) strategy that injects adversarial noises in the feature-level to smooth the decision boundary, aiming at for further utilizing the rest unselected but informative data. Although FAT is originally designed for the rest data, it can also be applied to the selected pseudo-labeled data. Based on the technics above, our PEFAT successfully boosts the classification performance in MIA from the point of trustworthy pseudo-labeled data selection and adversarial consistency regularization.

To summarize, our main contributions are three-fold. (1) Different from previous works that select pseudo-labeled data with a probability threshold, we present a new selection approach from the perspective of the loss distribution, which exhibits superior ability in high-quality pseudo-labeled data collection. (2) We propose a new adversarial regularization strategy to fully leverage the rest unlabeled but informative data, which benefits the model in decision boundary smoothing and better representation learning. (3) Extensive experimental results on three public medical image datasets and two natural image datasets demonstrate the superiority of the proposed PEFAT, which significantly surpasses other advanced SSL methods.

2. Related Work

Semi-supervised Learning. The paradigm of Semi-supervised Learning (SSL) can be concluded as learning from both labeled and unlabeled data, in the scenario of unlabeled data are the majority. Recently, various methods have been proposed, which can be roughly divided into three categories: pseudo-labeling [14, 22, 29, 37, 38], consistency regularization [12, 24, 25] and the combination of the above two [20, 32, 33, 41].

Pseudo-labeling. Pseudo-labeling-based methods follow the procedure of assigning pseudo-labels to unlabeled data vis a fixed or dynamic threshold, and then combine the manually annotated data for further re-training. For instance, ACPL [22] improves the accuracy of pseudo-labels by ensembling classifiers, and trains the model in an anti-curriculum manner. BoostMIS [40] takes the learning ability of model in different training stages into consideration, and proposes adaptive pseudo-labeling strategy for unlabeled data selection. Noise Student [38] tries to learn from training data iteratively, which generates pseudo-labels by the updated teacher network and redirects student network to learn from the whole data.

Unlike the aforementioned methods, we find it difficult to consider the probability threshold as a reliable standard for selecting clean pseudo-labeled data, since unlabeled data with both correct and incorrect pseudo-labels have similar probability distributions. However, noisy data typically result in high loss during training while clean data has the opposite effect. With this understanding, we suggest using pseudo-loss estimation to select pseudo-labeled data instead.

Consistency regularization. The core idea of consistency regularization is to minimize the output discrepancy for different views of unlabeled data, when adding different kinds of perturbations, $i.e.$, data augmentation and adversarial noises. SRC-MT [24] provides a data-relation consistency-based paradigm via self-ensembling learning. VAT [25] introduces virtual adversarial perturbations, which aims to regularize the predicted outputs by injecting the most adversarial noises in the image-level. AlphaMatch [8] proposes to use alpha-divergence and optimizes the model training in an EM-like fashion.

Compared to VAT [25], a method based on adversarial training, our proposed Feature Adversarial Training (FAT)

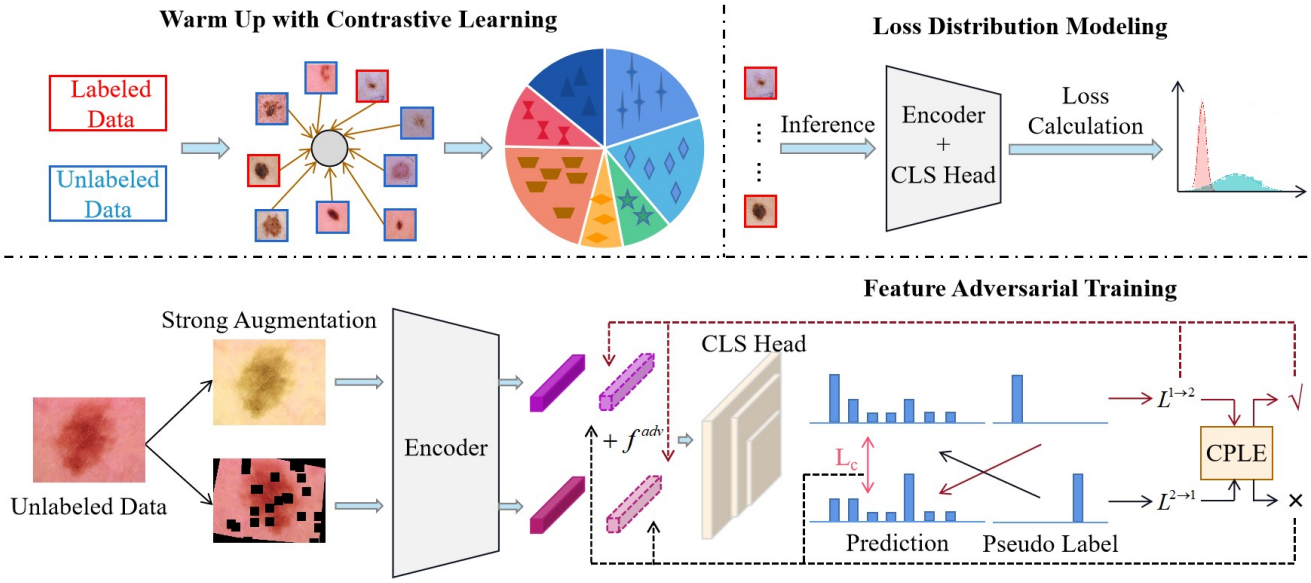


Figure 2. Illustration of our proposed PEFAT. We first warm up the model with contrastive learning on training data to learn unbiased representation. Then we set up a two-component GMM to construct the loss distribution calculated on labeled data. As for the unlabeled data utilization, we use the cross pseudo-loss estimation (CPLE) for trustworthy pseudo-labeled data exploration. Beyond that, adversarial noises are injected in the feature-level for better unlabeled data mining.

has two advantages: (1) globally, we inject the feature-level adversarial noises, which is more effective in discriminate informative mining, thereby can further improve the classification performance; and (2) detailly, the generation of adversarial noises is based on the output distribution between two different augmented views, which incorporates complementary information and produces less confirmation-biased adversarial noises.

3. Method

3.1. Preliminaries

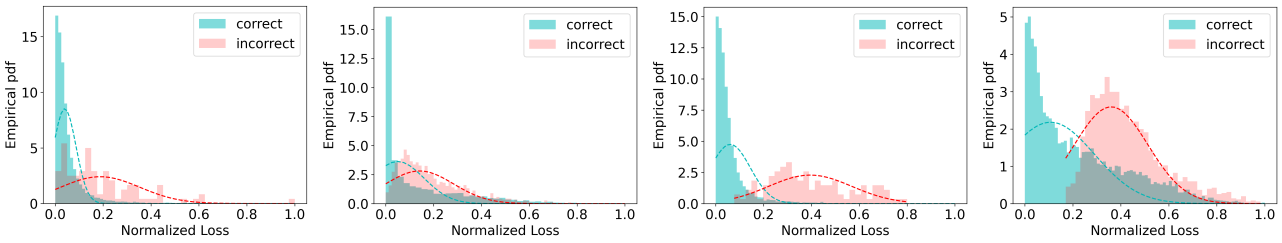
In the task of SSL classification, a labeled set $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$ and an unlabeled set $\mathcal{D}_u = \{(u_i)\}_{i=1}^{N_u}$ are commonly given, where N_l and N_u are number of samples with $N_l \ll N_u$. x_i is the input and $y_i = [y_i^1, y_i^2, ..., y_i^C] \subseteq \{0,1\}^C$ is the corresponding ground-truth with C class categories (note that more than one element in y_i can be non-zero in multi-label setting). Generally, we assume that the labeled and unlabeled data share the same distribution. The goal of this task is to establish an algorithm using both $\mathcal{D}_l$ and $\mathcal{D}_u$ tactfully. Normally, different methods differ in the usage of $\mathcal{D}_u$.

Figure 2 shows the workflow of PEFAT. We first warm up the classifier on the whole training data, and then model the loss distribution calculated on labeled images by GMM. As for the utilization of unlabeled data, we first find reliable pseudo-labeled data by feeding cross pseudo-loss to the fitted GMM. Beyond that, we establish an adversarial consistency regularization strategy by injecting feature-level adversarial noises to leverage the rest unselected but informative data. Although this strategy is initially proposed for the rest data, it also applies to the selected pseudo-labeled data. Below, Section 3.2 describes the details of loss distribution modeling, Section 3.3 shows the procedure of high-quality pseudo-labeled data selection, Section 3.4 provides the information of feature adversarial training for better unlabeled data learning and Section 3.5 summarizes the overall algorithm of PEFAT.

3.2. Loss Distribution Modeling

In pseudo-labeling-based SSL, three steps are commonly contained: (1) warm up a model h_θ parameterised by θ using $\mathcal{D}_l$; (2) generate pseudo-labels on $\mathcal{D}_u$ and collect high-confidence pseudo-labeled set $\tilde{\mathcal{D}}_u = \{(u_i, h_\theta(u_i))\}_{i=1}^{\tilde{N}_u}$; (3) re-train h_θ on the union of $\mathcal{D}_l \cup \tilde{\mathcal{D}}_u$. However, this paradigm has some limitations, as it highly relies on the model initialization on $\mathcal{D}_l$ and pseudo-labeled data in $\tilde{\mathcal{D}}_u$. And in most cases, the warm-upped h_θ will show confirmation bias [2] due to the unbalanced/partial distribution on $\mathcal{D}_l$, along with a large number of wrongly pseudo-labeled data for re-training.

Warm Up with Contrastive Learning. Inspired from previous works in self-supervised learning [4, 5], the contrastive loss is a useful technic to learn category-agnostic representation via maximizing the feature discrepancy among different views of a certain sample and the other



(a) (b) (c) (d)

Figure 3. Empirical probability density function (PDF) of the fitted GMM for loss distribution. (a) Training with FixMatch and loss distribution on labeled data; (b) Training with FixMatch and loss distribution on validation data; (c) Training with PEFAT and loss distribution on labeled data; (d) Training with PEFAT and loss distribution on validation data; (a) and (b) show zero-biased loss distribution, which is mainly attributed to over-confident prediction, while (c) and (d) present dividable distribution for pseudo-labeled data with correct and incorrect pseudo-labels, validating the effectiveness of cross pseudo-loss estimation.

samples. And in this work, we adopt the InfoNCE loss [26] to help h_θ learn more universal representation on $\mathcal{D}_l \cup \mathcal{D}_u$, instead of only focusing on the $\mathcal{D}_l$. The warm-upped loss is the following sum of InfoNCE loss $\mathcal{L}_{ct}$ on the whole training data and cross-entropy loss L_{ce} on the labeled data, defined as:

$$\mathcal{L}_{ct} = -\frac{1}{2|B|}\sum_{i=1}^{2|B|} log\frac{\exp(z_i \cdot z_i^+/\tau)}{\sum_{j=1}^{2|B|}\mathbb{1}_{(j\neq i)}\exp(z_i \cdot z_j/\tau)} \quad (1)$$

$$\mathcal{L}_{ce} = -\frac{1}{|B_l|}\sum_{i=1}^{|B_l|} D_{ce}(y_i, h_\theta(\hat{y}_i|A_w(x_i))) \quad (2)$$

where $|B| = |B_l| + |B_u|$, $|B_l|$ and $|B_u|$ denote the number of labeled and unlabeled samples in a mini-batch. $z_{i(j)}$ are the normalized feature embeddings and z_i^+ is the positive representation corresponding to z_i. τ is a temperature parameter. $D_{ce}(\cdot,\cdot)$ stands for the cross-entropy calculation. $\hat{y}_i$ and A_w represent predicted label and weak augmentation, respectively.

Loss Distribution Modeling on D_l. As indicated in Figure 1b, it is hard to regard the predicted probability as threshold to collect a clean pseudo-labeled set $\tilde{\mathcal{D}}_u$, due to the similar probability distribution for unlabeled samples with correct and incorrect pseudo-labels. Alternatively, wrongly pseudo-labeled samples tend to have a higher loss during the early training, which makes it possible to distinguish correct and incorrect samples by loss distribution (see Figure 3c). Based on the above observation, we assume that the overall loss distribution is composed of two normal distributions and further utilize the Gaussian Mixture Model (GMM) to fit the loss distribution on D_l. Formally, the instance-wise loss and probability density function (pdf) of GMM on loss ℓ_i can be formulated as:

$$\mathcal{L}(\mathcal{D}_l|h_\theta) = \{-y_i \log(h_\theta(\hat{y}_i|x_i)), x_i \in \mathcal{D}_l\} \quad (3)$$

$$\mathcal{I}(\ell_i) = \sum_{k=0}^{K-1} \pi_k I_k(\ell_i|\mu_k, \Sigma_k), \ell_i \in \mathcal{L}(\mathcal{D}_l|h_\theta) \quad (4)$$

where $\mathcal{L}(\mathcal{D}_l|h_\theta)$ is the set of loss on $\mathcal{D}_l$. $\pi_k \geq 0$ is the weight of the k-th gaussian component and $\sum_{k=0}^{K-1}\pi_k = 1$. For a certain loss ℓ_i, $\pi_k I_k$ indicates the probability of ℓ_i belonging to the k-th gaussian componet. After that, we use the Expectation Maximization (EM) algorithm to fit the GMM with the loss observation on $\mathcal{D}_l$, and the optimization procedure is maximizing the log-likelihood, which can be written as:

$$\hat{\theta}_{GMM} = \arg\max_{\theta_{GMM}}[\log \prod_{i=1}^{N_l} \mathcal{I}(\ell_i|\theta_{GMM})] \quad (5)$$

where $\theta_{GMM} = \{\pi_k, \mu_k, \Sigma_k\}$, $0 \leq k \leq K-1$. Based on the above process, GMM perceives the prior loss distribution on D_l, and is able to distinguish trustworthy pseudo-labeled samples by pseudo-loss distribution.

3.3. Trustworthy Pseudo-labeled Data Selection

Cross Pseudo-loss Estimation on D_u. Considering neural networks are generally over-confident to their predictions, we regard the prediction of one augmented view as the pseudo-label for the other augmented view, in order to avoid the zero-biased pseudo-loss distribution when treating unlabeled samples (see Figure 3b). In general, the cross prediction can be expressed as:

$$\hat{y}_i^{1\to 2} = \arg\max(h_\theta(A_{s1}(u_i))) \quad (6)$$

$$\hat{y}_i^{2\to 1} = \arg\max(h_\theta(A_{s2}(u_i))) \quad (7)$$

where A_{s1} and A_{s2} are two different strong augmentations, A_{s1} contains affine transformation, rotation and cutout, while A_{s2} contains grayscale, colorjitter and blur. $\hat{y}_i^{1\to 2}$

and $\hat{y}_i^{2\to1}$ are pseudo labels for views augmented from A_{s2} and A_{s1}, respectively. Finally, the cross pseudo-loss can be calculated as:

$$\ell_i^{1\to2} = -\hat{y}_i^{1\to2}\log(h_\theta(A_{s2}(u_i))) \tag{8}$$

$$\ell_i^{2\to1} = -\hat{y}_i^{2\to1}\log(h_\theta(A_{s1}(u_i))) \tag{9}$$

where $\ell_i^{1\to2}$ and $\ell_i^{2\to1}$ are the cross pseudo-loss.

Pseudo-labeled Sample Selection. Based on $\ell_i^{1\to2}$, $\ell_i^{2\to1}$ and the fitted GMM, we can select trustworthy pseudo-labeled sample u_i by the posterior probability, which can be formulated as:

$$p_{gmm} = \mathcal{I}(I_k|(\eta\ell_i^{1\to2} + (1-\eta)\ell_i^{2\to1})) \tag{10}$$

where $k = 0(1)$ stands for correct (incorrect) pseudo-loss component, η is a hyper-parameter to balance the weight between two pseudo-losses, p_{gmm} means the posterior probability of GMM.

In summary, to select unlabeled samples with correct pseudo-labels, we first calculate the instance-wise loss on D_l, and simulate the loss distribution by a two-component GMM. Finally, we select high-quality pseudo-labeled samples by the posterior probability of GMM, along with the cross pseudo-loss estimation scheme.

3.4. Feature Adversarial Training

Although we can effectively collect an almost clean pseudo-labeled set $\tilde{\mathcal{D}}_u$ by the Cross Pseudo-loss Estimation (CPLE), the rest of unselected data in $\overline{\mathcal{D}}_u = \mathcal{D}_u/\tilde{\mathcal{D}}_u$ are also informative for SSL training. Inspired by adversarial training [9, 25], adding adversarial perturbation is beneficial for smoothing decision boundaries, which is a practical strategy, especially in tackling edge-distributed samples. In this work, we propose Feature Adversarial Training (FAT), which injects targeted adversarial noises at the feature-level, aiming to explore information from unlabeled samples effectively. Specifically, given a sample u_i in $\tilde{\mathcal{D}}_u \cup \overline{\mathcal{D}}_u$, the generation of targeted adversarial noises (r_{i1}^{adv}, r_{i2}^{adv}) for two augmentation take the following format:

$$\begin{aligned} r_{i1}^{adv}, r_{i2}^{adv} = \arg\max_{\Delta r_1, \Delta r_2} [&J(h_\theta(p_i|z_i + \Delta r_1), \\ &h_\theta(p_i^+|z_i^+ + \Delta r_2))] \end{aligned} \tag{11}$$

where J is the Kullback-Leibler Divergence when u_i is from $\overline{\mathcal{D}}_u$. Otherwise, J is the cross-entropy loss since we can use the corresponding pseudo-label to replace $h_\theta(p_i|z_i + \Delta r_1)$. $||\Delta r_1|| \le \varepsilon$ and $||\Delta r_2|| \le \varepsilon$ are two random noises, ε is a hyper-parameter to regularize the applied noises. p_i and p_i^+ are the model predicted probability. However, we cannot acquire r_{i1}^{adv}, r_{i2}^{adv} according to Eq. (11) directly. To this end, [25] utilized the finite difference and power iteration to solve the problem, which can be simply performed using the following equation:

$$r_{i1}^{adv}, r_{i2}^{adv} \leftarrow \overline{\nabla_{\Delta r_1, \Delta r_2} J(h_\theta, z_i, z_i^+, \Delta r_1, \Delta r_2)} \tag{12}$$

Once r_{i1}^{adv} and r_{i2}^{adv} are attained, we utilize the feature adversarial training loss L_{FAT} for model optimization:

$$L_{FAT} = J(h_\theta(p_i|z_i + r_{i1}^{adv}), h_\theta(p_i^+|z_i^+ + r_{i2}^{adv})) \tag{13}$$

In summary, FAT seeks the direction of perturbation which can effectively alter the distribution at the feature-level, and thus benefits the model in robust learning and discriminative information mining.

Algorithm 1: PEFAT Algorithm

Input: Labeled dataset D_l; unlabeled dataset D_u; initialized model h_θ.

```
Initialize a two-componet GMM;
Warm up h_θ with Eq. (1) and Eq. (2);
for (x_i, y_i) ∈ D_l do
    Calculate loss l_{x_i} according to Eq. (3);
end
Fit GMM with {l_{x_i}}_{i=1}^{N_l} with Eq. (4) and Eq. (5);
for u_i ∈ D_u do
    Make cross prediction by Eq. (6) and Eq. (7);
    Get cross pseudo-loss by Eq. (8) and Eq. (9);
    Obtain p_gmm according to Eq. (10);
    if p_gmm^{k=0} > p_gmm^{k=1} then
        Calculate L_FAT and L_ce with pseudo-label;
    else
        Calculate L_FAT with Eq. (13);
    end
end
Return h_θ;
```

3.5. Overall Algorithm

We present the summary of PEFAT in Algorithm 1, which can be concluded as (1) finding out high-quality pseudo-labeled data from the perspective of the loss distribution, aiming at effectively mitigating the difficulty of dividing correct and incorrect pseudo-labeled data. Details include loss distribution modeling and cross pseudo-loss estimation; (2) adequately utilize the low confident unlabeled data. Here we try to make the two augmented views consistent by injecting feature-level adversarial noises, which can assist the model in having a better realization of these unlabeled data when the corresponding pseudo-label is unreliable. By incorporating the above two ideas, PEFAT successfully boosts the performance of medical image classification under the SSL setting.

Table 1. Performance comparison with other state-of-the-art SSL methods on NCT-CRC-HE dataset. "SENS", "PREC" and "ACC" are Sensitivity, Precision and Accuracy, respectively. We list the evaluation metrics when 100 and 200 labeled data are given. Best and second best results are shown in **bold** and underline, respectively.

Method	NCT-CRC-HE (200 labeled data)					NCT-CRC-HE (100 labeled data)				
	AUC	SENS	PREC	ACC	F1	AUC	SENS	PREC	ACC	F1
Baseline	97.86	78.12	83.06	80.63	76.31	96.48	73.85	76.25	73.29	73.48
MT [33]	98.07	81.89	83.91	81.55	81.19	97.15	77.51	78.81	77.97	77.07
FixMatch [32]	98.43	85.03	84.75	84.81	84.66	97.91	80.59	81.78	80.47	80.28
SimPLE [14]	98.57	85.80	85.56	85.59	85.48	98.01	83.37	83.46	82.72	82.91
CoMatch [20]	98.83	87.94	88.70	86.48	86.24	98.00	84.72	84.58	83.93	84.11
SimMatch [41]	99.02	88.19	88.36	88.31	87.98	98.03	85.07	84.50	84.24	84.43
Ours	**99.08**	**89.68**	**91.18**	**90.29**	**90.12**	**98.25**	**86.82**	**86.78**	**86.01**	**86.33**

Table 2. Performance comparison on ISIC2018 dataset. "SENS", "SPCE" and "ACC" stand for Sensitivity, Specificity and Accuracy, respectively. Evaluation metrics are reported with the percentage of 5% and 20% labeled data. Best and second best results are shown in **bold** and underline, respectively.

Method	ISIC2018 (20% labeled data)					ISIC2018 (5% labeled data)				
	AUC	SENS	SPEC	ACC	F1	AUC	SENS	SPEC	ACC	F1
Baseline	90.90	69.37	91.77	91.42	51.89	84.28	56.32	87.53	85.36	40.96
SRC-MT [24]	93.58	71.47	92.72	92.54	60.68	87.61	62.04	89.36	88.77	46.26
DS^{3}L [11]	93.85	70.33	92.29	92.53	61.08	85.08	58.82	89.52	89.27	44.19
ACPL [22]	94.36	72.14	-	-	62.23	-	-	-	-	-
RAC-MT [12]	94.42	73.41	92.68	93.27	63.95	87.92	59.34	90.51	91.11	48.54
Ours	**94.87**	**76.72**	**93.45**	**93.68**	**66.15**	**88.64**	**64.10**	**91.25**	**91.81**	**50.96**

4. Experiments

4.1. Setup

Datasets. We evaluate our method on three public medical image classification datasets, including **NCT-CRC-HE** [16], **ISIC2018** [7] and **Chest X-Ray14** [36]. Specifically, **NCT-CRC-HE** contains 100,000 colorectal cancer histology slides with nine categories, forming a multi-class classification task. We split the dataset into 70%/10%/20% for training/validation/test. And five evaluation metrics are reported: area under the ROC curve (AUC), Sensitivity, Accuracy, F1 score and Precision. For **ISIC2018**, it contains 10,015 skin lesion dermoscopy images with seven labels, which is also a multi-class dataset. We follow the same split as [12, 22, 24] for a fair comparison, which divides the entire dataset into 70%/10%/20% for training/validation/test. Evaluation metrics are the same as NCT-CRC-HE, except for replacing Precision with Specificity. **Chest X-Ray14** is a multi-label dataset with 112,120 chest x-rays from 30805 patients. It contains fourteen categories and each image may share multiple labels. To make a fair comparison with previous works, we adopt the same data split as [22–24], following 70%/10%/20% for training/validation/test. The evaluation metric of AUC is reported. Beyond that, we also conduct experiments on CIFAR-10 and CIFAR-100, which are presented in Supplementary Material.

Implementation Details. For all datasets, we use DenseNet-121 [15] as backbone with 224×224 input size. For model training, we use Adam optimizer [18] with a learning rate of 0.001. For a mini-batch, 16 labeled and 48 unlabeled images are contained. We train the model for 80 epochs, where 30 epochs and 50 epochs are respectively used to warm up and re-train the model. Hyper-parameter τ, ε and η are empirically set to 0.05, 1 and 0.5, respectively. All experiments are implemented in Pytorch [27] with two NVIDIA Gefore RTX 3080Ti GPUs.

4.2. Comparison with Existing Methods

Results on NCT-CRC-HE Dataset. In this part, we compare our method with five recently proposed SSL methods, including MT [33], FixMatch [32], SimPLE [14], CoMatch [20] and SimMatch [41]. All results are obtained using the same network architecture with the same input image size. As indicated by results in Table 1, we

can obtain the following findings: (1) our method continuously surpasses other SSL methods with the different number of annotated image data, $i.e.$, the performance gain of 1.49%∼7.79%, 2.48%∼7.27%, 1.98%∼8.74% and 2.14%∼8.93% in terms of sensitivity, precision, accuracy and f1 score, when given 200 labeled data; (2) compared to SimMatch, the most advanced SSL method that incorporates contrastive learning and consistency regularization, our method still achieves a slightly higher result (approximately 1.5% improvement) on the overall evaluation metrics, mainly owing to the CPLE for clean sample collection and FAT for fully leveraging the unlabeled samples; and (3) our method outperforms FixMatch by a large margin (roughly 5%∼6% performance gain in accuracy), further indicating the superiority of our proposed CPLE.

Results on ISIC2018 Dataset. Table 2 presents the results on ISIC2018 dataset, where competitive methods of SRC-MT [24], DS^{3}L [11], ACPL [22] and RAC-MT [12] are listed to compare. Our method again achieves the best results on all evaluation metrics with different label percentages ($i.e$, AUC:94.87%, sensitivity:76.72%, specificity:93.45%, accuracy:93.68%, f1 score:66.15%, in the setting of sharing 20% labeled data). This demonstrates that the intuition of filtering out wrongly predicted data by loss estimation is applicable in different datasets, as well as the benefits of FAT for learning discriminate information from unlabeled data.

Results on Chest X-Ray14 Dataset. Table 3 shows the results on Chest X-Ray14 dataset, where SOTA methods, $i.e.$, SRC-MT [24], S^{2}MTS^2 [23] and ACPL [22] are compared. Note that SRC-MT employees DenseNet-169 as the backbone with 384×384 input image size, while S^{2}MTS^2 and ACPL use DenseNet-121 as the backbone with 512×512 input size. Our PEFAT takes DenseNet-121 as the backbone with a smaller input size of 224×224. Again, PEFAT surpasses other competitive methods under various label percentage settings. Compared to pseudo-labeling methods UPS and ACPL, PEFAT outperforms them by 0.81%∼9.55%. Moreover, compared to the dual-path method NoTeacher, our approach consistently has approximately 3% performance gain, further vadilating the capability of PEFAT.

Table 3. Performance of mean AUC on Chest X-Ray14 dataset under the label percentage of 2%, 5%, 10%, 15% and 20%. Note that * denotes the methods employee DenseNet-169 as backbone with 384×384 input size, † means the methods use DenseNet-121 as backbone with 512×512 input size.

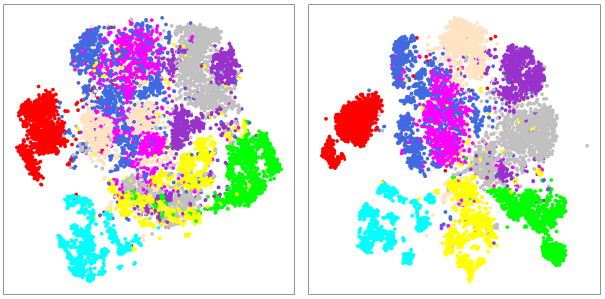
Method	Label Percentage				
	2%	5%	10%	15%	20%
Graph XNet* [3]	53.00	58.00	63.00	68.00	78.00
SRC-MT* [24]	66.95	72.29	75.28	77.76	79.23
UPS [29]	65.51	73.18	76.84	78.90	79.92
NoTeacher [34]	72.60	77.04	77.61	-	79.49
S^{2}MTS^{2†} [23]	74.69	78.96	79.90	80.31	81.06
ACPL† [22]	74.82	79.20	80.40	81.06	81.77
Ours	**75.06**	**79.54**	**80.93**	**81.56**	**82.58**

Table 4. Ablation study of each module in PEFAT on NCT-CRC-HE dataset. Results are reported in the case of 100 labeled data. $*$ and † denote singly employing FAT on the selected and unselected pseudo-labeled data, respectively.

Method	AUC	SENS	PREC	ACC	F1
Baseline	96.48	73.85	76.25	73.29	73.48
CPLE	98.09	84.57	83.89	84.16	84.65
CPLE+VAT	98.13	85.00	84.14	84.34	84.79
CPLE+FAT*	98.15	85.10	84.66	84.42	84.70
CPLE+FAT†	98.18	85.91	85.76	85.65	85.73
CPLE+FAT	**98.25**	**86.82**	**86.78**	**86.01**	**86.33**

4.3. In-Depth Analysis

Ablation study. Results of the ablation study are shown in Table 4. As we can see, the evaluation metrics increase significantly by simply utilizing CPLE (10.87% improvement in accuracy compared to baseline), indicating the superior of our proposed pseudo-labeled sample selection scheme. Beyond that, it is worth noting that CPLE surpasses other advanced pseudo-labeling-based methods, $i.e.$, FixMatch and SimPLE, further demonstrating the capability of CPLE. Row3 and Row6 show the availability of VAT and FAT, respectively. We can find that there is little performance gain (0.18% improvement in accuracy) when using VAT, mainly due to insufficient and biased adversarial noises. Nevertheless, FAT boosts the accuracy by 1.85%. Moreover, Row4 and Row5 present the effects of singly applying FAT on the selected and unselected pseudo-labeled data, respectively.



(a) (b)

Figure 4. The t-SNE visualization on NCT-CRC-HE validation set. (a) is the result when using VAT; (b) shows the feature embedding when using FAT.

Q1: What benefits can FAT bring? To further explore the effects of FAT, we respectively exhibit the t-SNE visualization results when employing VAT and FAT. As depicted in Figure 4a, although VAT shows the ability to smooth the decision boundary, feature embedding of several categories still mixes. This phenomenon indicates the limitation of VAT. In contrast, Figure 4b presents better clusters and more distinguishable decision boundary when applying FAT, demonstrating its effectiveness in handling boundary-distributed samples and learning dividable representation.

Q2: What is the relation between the model predicted probability and posterior probability of GMM for the pseudo-labeled data? Since we select pseudo-labeled data based on the posterior probability of the fitted GMM, it is interesting to discover its relation with the model predicted probability. As illustrated in Figure 5, we can observe that wrongly pseudo-labeled samples with high model predicted probability might also show low posterior probability, mainly caused by cross prediction, $i.e.$, disagreement for the prediction. Since we fit a two-component GMM according to the loss distribution, pseudo-labeled data with a posterior probability higher than 0.50 are selected by us. We can find that correct pseudo-labeled data generally have model predicted probability ranging from 0.70 to 0.95. Our proposed CPLE can effectively maintain the unlabeled data with correct pseudo-labels, while traditional pseudo-labeling-based methods will fall into a dilemma. That is, increasing the threshold will greatly reduce the number of correct samples, conversely decreasing the threshold will introduce more incorrect samples. Besides, here we list some quantitative results, 1466 correct and 132 incorrect pseudo-labeled data lie on the right of the dotted line, while the other side contains 129 correct and 495 incorrect data. These results validate the convenience of CPLE, which can collect a high-quality pseudo-labeled set without the trade-off between threshold and accuracy.

Q3: To what extent can we trust the selected pseudo-labeled data? To answer this question, we conduct an experiment to compare with probability-based pseudo-labeling strategy (including thresholds of 0.80, 0.85, 0.90, 0.95). From results in Table 5, we can draw the following conclusions: (1) the higher the probability threshold, the higher the proportion of correctly predicted samples; (2) although the ratio of correct pseudo-labeled samples is high with a higher threshold, numerous unlabeled labeled are abandoned; and (3) compared to traditional pseudo-labeling strategy, the superior of CPLE lies in maintaining a considerable number of correctly predicted samples as well as minimizing the error rate. For instance, compared to $\delta = 0.80$, CPLE has almost two times correctly predicted samples. And compared to $\delta = 0.95$, although CPLE selects more incorrect pseudo-labeled samples, the error rate is still lower. Beyond that, it should be noted that clean samples discovered by CPLE are approximately six times than $\delta = 0.95$. To summarize, we can select the largest pseudo-labeled set with the lowest error rate by leveraging CPLE.

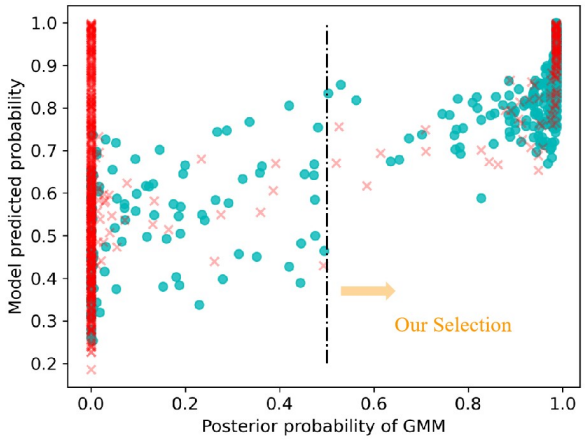


Figure 5. The relation between model predicted probability and posterior probability of GMM. Red × and Turquoise • denote unlabeled data with incorrect and correct pseudo-labels.

Table 5. Experiments conducted on NCT-CRC-HE validation set. $\delta = K$ means using probability threshold K to select pseudo-labeled samples. Ratio = Correct / Selected.

Method	Selected		Unselected ↓	Ratio ↑
	Correct ↑	Incorrect ↓		
δ=0.80	3821	531	5648	87.80
δ=0.85	3172	394	6434	88.95
δ=0.90	2387	257	7356	90.28
δ=0.95	1184	**112**	8704	91.36
CPLE	**6490**	592	**2918**	**91.64**

5. Conclusion

In this paper, we propose a new method, PEFAT, for semi-supervised medical image classification, stemming from the point of pseudo-loss estimation and adversarial training. PEFAT can effectively judge the quality of pseudo-labels, and thus directly benefit the model by learning from reliable pseudo-labeled data. Moreover, we also introduce an adversarial-based consistency regularization strategy for sufficiently leveraging the unselected but informative data. Extensive experiments on three medical and two natural image datasets demonstrate the superior of PEFAT as well as its versatility in various settings.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *ICML*, pages 312–321, 2019. 2

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*, pages 1–8, 2020. 3

[3] Angelica I Aviles-Rivero, Nicolas Papadakis, Ruoteng Li, Philip Sellars, Qingnan Fan, Robby T Tan, and Carola-Bibiane Schönlieb. Graphxnet chest x-ray classification under extreme minimal supervision. In *MICCAI*, pages 504–512, 2019. 7

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 1, 3

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 1, 3

[6] Yanbei Chen, Massimiliano Mancini, Xiatian Zhu, and Zeynep Akata. Semi-supervised and unsupervised deep visual learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[7] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 6

[8] Chengyue Gong, Dilin Wang, and Qiang Liu. Alphamatch: Improving consistency for semi-supervised learning with alpha-divergence. In *CVPR*, pages 13683–13692, 2021. 2

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2014. 2, 5

[10] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021. 1

[11] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *ICML*, pages 3897–3906, 2020. 6, 7

[12] Wenlong Hang, Yecheng Huang, Shuang Liang, Baiying Lei, Kup-Sze Choi, and Jing Qin. Reliability-aware contrastive self-ensembling for semi-supervised medical image classification. In *MICCAI*, pages 754–763, 2022. 2, 6, 7

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

[14] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. Simple: similar pseudo label exploitation for semi-supervised classification. In *ICCV*, pages 15099–15108, 2021. 2, 6

[15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 1, 6

[16] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Medicine*, 16(1):e1002730, 2019. 6

[17] Youngwook Kim, Jae Myung Kim, Zeynep Akata, and Jungwoo Lee. Large loss matters in weakly supervised multi-label classification. In *CVPR*, pages 14156–14165, 2022. 2

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[20] Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *ICCV*, pages 9475–9484, 2021. 2, 6

[21] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. 1

[22] Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. In *CVPR*, pages 20697–20706, 2022. 2, 6, 7

[23] Fengbei Liu, Yu Tian, Filipe R Cordeiro, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Self-supervised mean teacher for semi-supervised chest x-ray classification. In *International Workshop on Machine Learning in Medical Imaging*, pages 426–436, 2021. 6, 7

[24] Quande Liu, Lequan Yu, Luyang Luo, Qi Dou, and Pheng Ann Heng. Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE Transactions on Medical Imaging*, 39(11):3429–3440, 2020. 2, 6, 7

[25] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018. 2, 5

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019. 6

[28] Haim Permuter, Joseph Francos, and Ian Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, 39(4):695–706, 2006. 2

[29] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An

uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *ICLR*, 2020. 2, 7

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 1

[31] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221, 2017. 1

[32] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, pages 596–608, 2020. 2, 6

[33] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pages 1195–1204, 2017. 2, 6

[34] Balagopal Unnikrishnan, Cuong Manh Nguyen, Shafa Balaram, Chuan Sheng Foo, and Pavitra Krishnaswamy. Semi-supervised classification of diagnostic radiographs with noteacher: a teacher that is not mean. In *MICCAI*, pages 624–634, 2020. 7

[35] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Transactions on Medical Imaging*, 37(7):1562–1573, 2018. 1

[36] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, pages 2097–2106, 2017. 6

[37] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *CVPR*, pages 4248–4257, 2022. 2

[38] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10687–10698, 2020. 2

[39] Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In *ECCV*, pages 558–575, 2022. 1

[40] Wenqiao Zhang, Lei Zhu, James Hallinan, Shengyu Zhang, Andrew Makmur, Qingpeng Cai, and Beng Chin Ooi. Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *CVPR*, pages 20666–20676, 2022. 2

[41] Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *CVPR*, pages 14471–14481, 2022. 2, 6