A Self-improving Skin Lesions Diagnosis Framework Via Pseudo-labeling and Self-distillation

Shaochang Deng[1] SHAOCHANGD@FOXMAIL.COM
Mengxiao Yin[1,2] YMX@GXU.EDU.CN
Feng Yang[1,2,*] YF@GXU.EDU.CN
1 School of Computer and Electronics and Information, Guangxi University, Nanning, Guangxi 530004, PR China

2 Guangxi Key Laboratory of Multimedia Communications Network Technology, Guangxi University, Nanning, Guangxi 530004, PR China

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

In the past few years, supervised-based deep learning methods has yielded good results in skin lesions diagnosis tasks. Unfortunately, obtaining large of labels for medical images is expensive and time consuming. In this paper, we propose a self-improving skin lesions diagnosis (SISLD) framework to explore useful information in unlabeled data. We first propose a semi-supervised model f, which combining consistency and class-balanced pseudo-labeling to make full use of unlabeled data in scenarios with sparse manually labeled samples, and obtain a teacher model f_t by semi-supervised self-training. Then, we introduce self-distillation method to enable knowledge distillation for the diagnosis of skin lesions. Finally, we measure diagnostic effectiveness in the context of label sparsity and class imbalance. The experiments on skin lesion images dataset ISIC2018 shows that SISLD achieves significant improvements in AUC, Accuracy, Specificity and Sensitivity.

Keywords: Skin lesion images; Diagnosis; Consistency; Pseudo-labeling; Self-distillation;

1. Introduction

The skin is the largest organ of the human body. Sun's UV-radiation, smoking, alcohol, and other internal and external factors can cause visible and touchable skin lesions. Skin lesions primarily include melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AKIEC), benign keratosis (BKL), dermatofibroma (DF), and vascular lesions (VASC), which can be divided into benign and malignant lesions. Malignant lesions may lead to skin cancer and metastasize to other organs and tissues(Burdick et al., 2018), while some malignant lesions are skin cancers. Skin cancer is a public health problem that cannot be ignored, with melanoma being the deadliest skin cancer, capable of spreading to different parts and organs of the human body. In 2020, new cases of global melanoma were estimated to exceed 324,000, with almost 57,000 deaths(Sung et al., 2021). Some types of lesions may not be as deadly as melanoma, such as basal cell carcinoma, but these lesions can spread and cause disfigurement and are life-threatening(Rubin et al., 2005). Therefore, the current diagnosis of skin lesions at an early stage is particularly important. If diagnosed

*Corresponding author: yf@gxu.edu.cn

early, skin cancer can be cured by excision. The huge difference in skin appearance, the similarities of different skin lesions, and intra-class variation make it difficult to compare skin lesions and typical tissues with the naked eye, which brings difficulties to the early and correct diagnosis of skin lesions. Accurate diagnosis of skin lesions is a big challenge in medical image diagnosis.

Dermoscopy, the primary tool for skin lesion diagnosis(Sultana et al., 2012), can eliminate skin surface reflections, and magnify the lesion area to enhance the visualization of deeper skin. Although it can bring richer visual information than the naked eye, the visual inspection of skin lesions is still related to subjective factors such as experience and is time consuming and error prone. Even experienced dermatologists guided by well-established methods such as the 7-point checklist(Kawahara et al., 2018)and the ABCDE rules(Goldsmith and Solomon, 2007) may give a dissimilar diagnosis. Therefore, a computer-aided diagnosis system capable of analyzing dermoscopy images can be established to assist dermatologists in their diagnosis.

In recent years, deep learning has developed into the current state-of-the-art classification algorithms, including supervised learning and semi-supervised learning, which are generally employed in various image classification and recognition tasks (Szegedy et al., 2015; He et al., 2016). Supervised learning has become an essential part of many advanced Skin lesion diagnosis approaches. For example, Kassani and Kassani (2019) employ deep learning models such as Resnet50, AlexNet(Krizhevsky et al., 2012), etc. to diagnose skin lesions. They alleviate the negative impact of class imbalance through data augmentation. Wang et al. (2021) proposed a multi-level attention learning network to improve the diagnosis of melanoma. They designed a local learning branch with a Skin Lesion Localization module to help the network learn features from regions of interest. Moreover, they improved feature recognition ability by combining information from global and local branches through a weighted feature integration module. The success of supervised deep learning methods comes from a large amount of labeled data. For most image-based diseases diagnosis tasks, it is arduous to obtain a large number of labels. While the unsupervised methods(Chen et al., 2020; He et al., 2020) does not required labels and focus on the samples themselves, they are not effective in the diagnosis task. As a compromise, semi-supervised learning (SSL) utilize both labeled and unlabeled data to train the model, which is more suitable for image-based diseases diagnosis. The consistency-based semi-supervised approach has shown its effectiveness. The Temporal Ensembling(Laine and Aila, 2016) maintains a huge exponential moving average (EMA) matrix during training and adopts it as consistency target. Maintaining a huge matrix during training will significantly increase training time. The Mean Teacher(Tarvainen and Valpola, 2017) framework updates the weights of the teacher model by the EMA weights of the student model, avoiding the need to maintain a huge matrix during training. Liu et al. (2020) introduced a Sample Relational Consistency paradigm on the Mean Teacher framework, modeled the relational information of different samples through the Gram matrix, and effectively used unlabeled data by minimizing the SRC paradigm. However, the SRC-MT completely ignores the obtained model, which include valuable knowledge. The performance of skin lesions diagnosis is limited by the complexity of the lesions themselves, also related to the imbalanced of skin lesions images. The excessive imbalance of the images of skin lesions may lead to poor performance. To develop a high-performance CAD system capable of diagnosing skin lesions, this work explores the diagnosis in the context of label scarcity and class imbalance. The main contributions of the proposed method are summarized as follows:

1. We propose a semi-supervised framework SISLD for skin lesion classification. The pseudo-labeling process of SISLD combine the consistency regularization and class-balanced pseudo-labeling. We leverage the labeled and unlabeled images to alleviate the interference of labeled samples scarce and class imbalance.

2. To strengthen the performance of model, we introduce self-distillation in distillation process.

3. We have carried out several extensive experiments to verify the validity of our framework on ISIC2018 dataset.

2. Method

In this section, we show the mechanism of our framework. To address the issues of class imbalance and label scarcity in skin lesion images diagnosis task, we aim to utilize unlabeled data via joint consistency regularization and class-balanced pseudo-labeling. Furthermore, without requiring additional labels, we improve the model performance on imbalanced datasets through pseudo-labels supervised self-distillation. The proposed SISLD framework as shown in Fig.1 consists of two parts: a consistent and class-balanced pseudo-labeling framework in the upper half and a pseudo-labels supervised self-distillation framework in the lower half. The training algorithm for SISLD is shown in Algorithm 1.

2.1. SISLD Framework

The training process of SISLD include pseudo-labeling and self-distillation.SISLD framework consists of four models, student model f_s and teacher model f_t of pseudo-labeling process, student model g_s and teacher model g_t of self-distillation process. Let the labeled set be represented as $D_L = \{(x_i, y_i)\}_{i=1}^{N}$, and the unlabeled set as $D_U = \{(x_i)\}_{i=1}^{M}$, where x_i is the input 2D skin lesion image, y_i is the corresponding One-Hot ground truth, y_i is the pseudo-labels. In pseudo-labeling process, we train f_s and f_t on all data and use the following minimized combined loss function to optimize the network:

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}_s \left(f_s\left(x_i;\theta\right), y_i, \widehat{y_i}\right) + \mathcal{L}_u \left(\{x_i\}_{i=N+1}^{N+M}; f_s, f_t, \theta, \eta, \theta', \eta'\right) \tag{1}$$

Where θ and and θ' are the parameters of the student and teacher model respectively. η and η' donate the perturbations of the input image. During the training process, θ update by optimizer, and θ' update by Exponential Moving Average (EMA). Specifically, we update θ' at the training step t: $\theta'_t = \lambda\theta'_{t-1} + (1-\lambda)\theta_t$, λ denote the coefficient that control the weight of the EMA. The training detail of pseudo-labeling process will be showed in section 2.2.

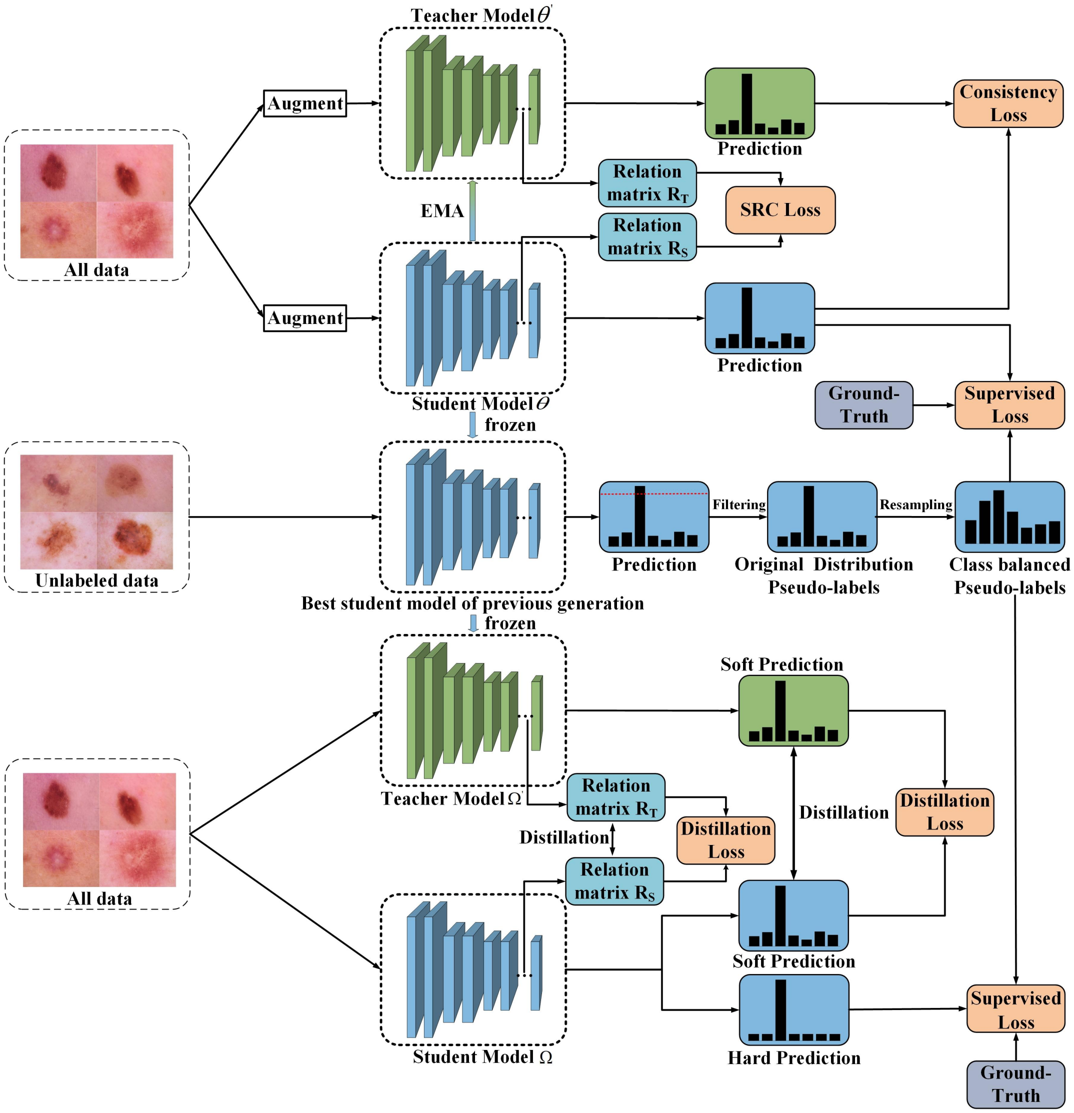


Figure 1: An overview of SISLD framework. SISLD consists of two pairs of student-teacher models, corresponding to two processes. All models are based on Densenet121 networks. In the pseudo-labeling process (top), our framework optimizes the student model f_s by total loss, and teacher model optimize by EMA. The pseudo-labels come from the prediction of student model f_s, filter pseudo-labels and change its distributions to alleviate class imbalance and noisy pseudo-labels(mid). As for self-distillation (bottom), SISLD initialize teacher model g_t by the weights of student model f_s, and utilize g_t to guide the training of student model g_s.

Next, we perform self-distillation process to train student modelg_s and teacher model g_t. The total optimization objective of the self-distillation process can be formulated as following:

$$\min_{\Omega} \sum_{i=1}^{N} \mathcal{L}_s \left(g_s \left(x_i ; \Omega \right), y_i, \widehat{y_i} \right) + \mathcal{L}_{kd} \left(\{x_i\}_{i=N+1}^{N+M} ; g_s, g_t, \Omega, \Omega' \right) \tag{2}$$

Where Ω and Ω' are the parameters of the student and teacher model respectively. Different from the pseudo-labeling process, the parameters of two models all updates by optimizer. More detail of self-distillation process will be showed in section 2.3.

2.2. Class balanced pseudo-labeling

The key idea of pseudo-labeling is utilize the existing knowledge of the model to keep increasing the number of pseudo labels. Consistency regularization is that for any input, under different perturbations, the model can still produce the same output as the original. It should be noted that most SSL methods have examined only consistency regularization or pseudo-labeling. The different mechanisms make it possible to combine two strategies. Currently, a framework called FixMatch (Sohn et al., 2020) achieves state-of-the-art performance across many standard SSL baselines by jointing two SSL strategies. Inspired by FixMatch, we combine pseudo-labeling and consistency regularization on the SRC-MT framework. We first train SISLD on all data. Then utilize the optimal student model f_s to generate pseudo labels for unlabeled samples. Finally, start next generation and add pseudo labels to the labeled set.

One primary problem with pseudo-labeling is that since pseudo-labels cannot be guaranteed to be 100% correct, the model will inevitably be disturbed by noise labels. Multiple generations of training will amplify this interference, which is also the disadvantage of pseudo-labeling(Rizve et al., 2021). To alleviate the influence of noisy labels and obtain trustworthy pseudo-labels, we filter pseudo-labels using a threshold value. Specifically, for the model predictions, only predictions with confidence greater than τ are retained as pseudo-labels. To determine the threshold value, we did the following experiments and the final value of was determined to be 0.90. The ablation studies are shown in Section 3.7.

Despite the fact that the traditional pseudo-labeling strategy has remarkable success, it still has shortcomings when dealing with class-imbalance problems. Resampling is a common method to alleviate class imbalance. In previous studies, resampling was done on labeled samples. In skin lesion diagnosis, it is unacceptable to drop the labeled data. Consequently, we employ resampling on pseudo-labels. An observation of CReST (Wei et al., 2021) shows that in the case of class imbalance, the semi-supervised model has low precision for the majority class and high precision for the minority class, which means the pseudo-label of the minority class is trustworthy. Thus, the specific approach of the class-balanced strategy in this work is to reduce the number of pseudo-labels of the majority class and increase the quantity of pseudo-labels of minority classes. Reducing the quantity of majority classes pseudo-labels with low precision may decrease the probability of introducing noisy labels. Concretely, let Z_i represent the original distributions of the i-th class($i \in (0, 1, 2, 3, 4, 5, 6)$), P_n denotes the quantity of pseudo-labels, α is the balance factor. For a certain quantity of pseudo-labels P_n , the specific quantity of pseudo-labels of the i-th class($i \in (0,2,3,4,5,6)$) is C_i, and the formula for C_i is as follows:

$$C_i\left(P_n\right)=\left\{\begin{array}{l}Z_i+\left\lceil\alpha\left(\frac{P_n}{7}-Z_i\right)\right\rceil, i \neq 1 \\ \\ P_n-\sum_{i=(0,2,3,4,5,6)} C_i\left(P_n\right), i=1\end{array}\right. \tag{3}$$

In the pseudo-labeling process, we employ SRC-MT as basic framework. To optimize the student model f_s, we utilize the following minimized combined loss function $\mathcal{L}_{pl}$ to optimize the network:

$$\mathcal{L}_{pl}=\mathcal{L}_s+\mathcal{L}_u, \text{ with } \quad \mathcal{L}_u=\mathcal{L}_c+\mathcal{L}_{src} \tag{4}$$

In the Eq 4, $\mathcal{L}_s$ represent the supervised loss, which is calculated by ground-truth, pseudo-labels and the perdition of student model f_s. When training at first generation, $\mathcal{L}_s$ was calculated by ground-truth and the perdition of student model f_s. $\mathcal{L}_s$ represent the unsupervised consistency loss which consist of individual consistency and relation consistency. The individual consistency loss $\mathcal{L}_c$ is defined as following:

$$\mathcal{L}_c=\sum_{i=1}^{N+M} \mathbb{E}_{\eta^{\prime} \eta}, \left\|f_t\left(x_i, \theta^{\prime}, \eta^{\prime}\right)-f_s\left(x_i, \theta, \eta\right)\right\|_2^2 \tag{5}$$

Most of the studies of consistency regularization are however focused on single sample. Unfortunately, the intrinsic relationship between different samples contains rich semantic information. To utilize this semantic information, SRC paradigms establish the internal relationship model between samples through the Gram matrix. Specifically, for n input samples, let $F^l \in \mathbb{R}^{n \times CHW}$ represent the activation map of the output of penultimate layer, and reshape F^l as $A^l \in \mathbb{R}^{n \times CHW}$, where H and W are the spatial dimensions of the feature map, C is the quantity of channels, then calculate the matrix $G^l \in \mathbb{R}^{n \times n}$ as $G^l=A^l \cdot \left(A^l\right)^T$, and perform l2 normalization on each row G_i^l of G^l to obtained sample relation matrix R^l, which is expressed as:

$$R^l=\left[\frac{G_1^l}{\left\|G_1^l\right\|_2^2}, \ldots, \frac{G_n^l}{\left\|G_n^l\right\|_2^2}\right]^T \tag{6}$$

where G_{ij} is the inner product between the vectorized activation maps A_i^l and A_j^l, which means the similarity between the activation maps of the i-th sample and the j-th sample in a batch. Thus, SRC loss define as follows:

$$\mathcal{L}_{src}=\sum_{X \in\left(D_U U D_L\right)} \frac{1}{n}\left\|R^l(X ; \theta, \eta)-R^l\left(X ; \theta^{\prime}, \eta^{\prime}\right)\right\|_2^2 \tag{7}$$

$R^l\left(x_i ; \theta, \eta\right)$ and $R^l\left(x_i ; \theta^{\prime}, \eta^{\prime}\right)$ are different perturbations sample relation matrix computed on x_i. When a generation is completed, we utilize the optimal student model $\theta^{\prime}$ to generate pseudo labels for unlabeled samples, the distributions of pseudo labels was control by Eq 3. With each generation, the quantity of pseudo labels increases by 350(5%). Then, start next generation and add the pseudo labels to labeled set. Repeat the above process until the maximum number of generations. Since the completion of the pseudo-labeling process, we select and save the optimal student model for all generations.

2.3. Pseudo-labels Supervised Self-Distillation

Recent evidence suggests that simply average outputs of multiple different models is a productive way to boost the performance (Dietterich, 2000) . However, making predictions is inefficient and computationally expensive. For this reason, Hinton et al. (2015) propose the concept of knowledge distillation. Knowledge distillation can compress the model so that its parameters are significantly reduced and it has the performance of a large model. It utilizes knowledge distilled from a teacher model to guide the training of student model. Knowledge distillation is usually based on models trained from large datasets, yet the amount of data is often small for some specific tasks. To deal with this dilemma, Zhang et al. (2019) propose a self-distillation strategy. Self-distillation achieve performance improvements by employing itself as the teacher model and does not require additional data as well as calculated costs. In skin lesion diagnosis, even small performance improvements can save more patients' lives. To further boost the performance of our framework, we introduce self-distillation in it. In previous process, we obtain a best model f_s, which satisfies the teacher model required for self-distillation. The optimal student model f_s is used as the teacher model in the self-distillation process to guide the training of student model g_s in self-distillation process.

In self-distillation process, we train the models for 60 epochs. The parameters of student model Ω is initialized by the optimal student model f_s. Consider an image x_i randomly picked from $\{\mathrm{D}_U \cup \mathrm{D}_L\}$ with label y_i and pseudo labels $\widehat{y}_i$. Taking the image as input, the student model generates hard prediction P_h and soft prediction P_s, and the teacher model generates soft labels P_t. During training, the parameter of student model is optimized to minimize the soft label loss and the hard label loss. The $\mathcal{L}_{SD}$ formula is as follows:

$$\mathcal{L}_{sd} = \frac{1}{n} \sum_{i=1}^{n} T^2 \cdot D_{KL}\left(P_s \| P_t\right) \tag{8}$$

Where n is the quantity of samples and T is the temperature parameter. The $\mathcal{L}_{rd}$ formula is as follows:

$$\mathcal{L}_{rd} = \sum_{x_i \in (D_U \cup D_L)} \frac{1}{n} \left\| R^l(x_i; \Omega) - R^l\left(x_i; \Omega'\right) \right\|_2^2 \tag{9}$$

Overall, the training proceeds with simultaneous minimization of the overall loss $\mathcal{L}_{kd}$, which is the sum of the distilling loss $\mathcal{L}_{sd}$, $\mathcal{L}_{rd}$ and supervised loss $\mathcal{L}_s$:

$$\mathcal{L}_{kd} = \mathcal{L}_s + \mathcal{L}_{sd} + \mathcal{L}_{rd} \tag{10}$$

We employ cross-entropy loss as the hard label loss $\mathcal{L}_s$ is calculated by ground-truth, pseudo-labels and P_h.

Algorithm 1 Training Procedure of the Proposed Method

```
Input: $x_i \in D_L + D_U, y_i \in D_L$
Output: student model's parameter $\theta$ and teacher model's parameter $\theta^{'}$, student model's
    parameter $\Omega$ and teacher model's parameter $\Omega^{'}$
for $T$ in [1, numgenerations] do
    for $T$ in [1, numepochs] do
        sample batch $B = B_l + B_u$ , where $B_l = (x_i, y_i) \in D_l$ and $B_u = x_i \in D_u$
        computing student model's prediction $f_s\left(x_i;\theta,\eta\right)$ and teacher model's prediction
            $f_t\left(x_i;\theta',\eta'\right)$ , $i \in \{1, ..., n\}$ where $n$ is the batch size
        computing the pseudo-labeling total loss according to Eq 4
        update $\theta$ using optimizer,update $\theta^{'}$ by EMA
    end for
    using the optimal student model $f_t$ to generate pseudo labels $\widehat{y}_i$
    filtering and change the distribution of pseudo labels by Eq 3
end for
for $T$ in [1, numepochs] do
    sample batch $B = B_l + B_u$ , where $B_l = (x_i, y_i) \in D_l$ and $B_u = x_i \in D_u$
    computing student model's prediction $g_s\left(x_i;\Omega\right)$ and teacher model's prediction
        $g_t\left(x_i;\Omega^{'}\right)$ , $i \in \{1, ..., n\}$ where $n$ is the batch size
    computing the self-distillation total loss according to Eq 10
    update $\Omega$ and $\Omega^{'}$ using optimizer
end for
return $\theta$, $\theta^{'}$,$\Omega$ and $\Omega^{'}$
```

3. Experiments

3.1. Dataset

The dataset was provided by the International Skin Imaging Collaboration (ISIC) 2018 Classification Challenge. The ISIC2018 dataset include 10,015 dermoscopic images of 600×450 size. ISIC2018 has 7 types of lesions, namely MEL, NV, BCC, AKIEC, BKL, DF, VASC. These lesions have a highly inter-class similarity and significant intra-class variation. Since only the training set has ground-truth, we randomly divide the training set, 70% for training, 20% for testing, and 10% for validation. In a real application scenario, the distribution of the obtained data is imbalanced, so the distribution of the unknown data is also likely to be imbalanced. The data distribution is roughly similar to training set. The distribution of dataset is shown in Fig. 2.

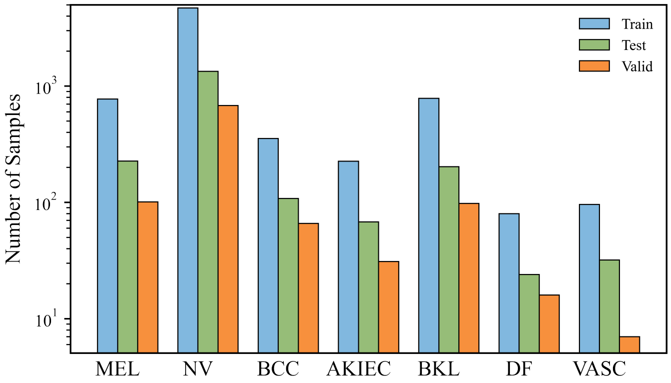


Figure 2: The distribution of the dataset.

3.2. Data Pre-Processing

Due to the uncertainty of the lesion area, skin lesion images may carry hair noise which block the original disease features. Meanwhile, hair may regard as disease features by the model, thus affecting the model training. Therefore, we utilize traditional image processing algorithms to address this issue. First, convert the original image to grayscale, employ Blackhat operation on grayscale images. Then, use threshold segmentation to get mask and utilize OpenCV's inpaint algorithm to repair images. The hair removal effect is shown in the Fig. 3.

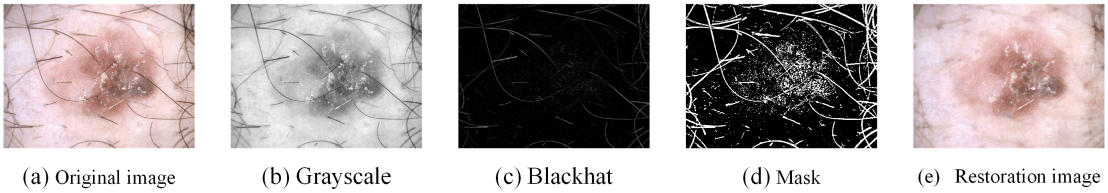

(a) Original image (b) Grayscale (c) Blackhat (d) Mask (e) Restoration image

Figure 3: The example of hair removal results.

3.3. Implementation Details

We implement the proposed framework using PyTorch framework, and implement our experiments on a Nvidia RTX 2060. In this work, we use the recurring SRC-MT framework as the baseline. As the quantity of images of some classes is rare, for all process, we employ different data augmentation techniques to expand the dataset, such as horizontal flipping, vertical flipping, Gaussian noise, etc. We employ Adam as the optimizer, the learning rate is $1e^{-4}$ and decayed with a power of 0.9 after each epoch, the dropout rate is 0.2. The batch size is set to 16, which is composed of 4 labeled samples and 12 unlabeled samples. We convert the input pixel range to 0-1, and resized images from 600 × 450 to 224×224. We also normalized the images by subtracting the proposed mean RGB values in the ImageNet dataset. EMA decay rate is set to 0.99. We employ random rotation, translation, and horizontal flipping as perturbations of input samples. The number of pixels for horizontal and vertical translation is in range -2% to 2% of the image width, and the probability of flipping is 50%. The random rotation ranges from -10 to 10. Specifically, in the first generation of pseudo-labeling process, the parameters of student model f_s is initialized by a pretrained model on ImageNet. The temperature of self-distillation process is set to 2.

3.4. Evaluation Metrics

The task of ISIC2018 is a multi-class classification problem, to quantitatively evaluate the proposed method, we employ Area Under receiver operating characteristic Curve (AUC), Accuracy, Sensitivity, Specificity as the evaluation criteria.

$$AUC = \int_0^1 t(f)d(f) \tag{11}$$

$$\text{Accuracy } = \frac{TP + TN}{TP + TN + FP + FP} \tag{12}$$

$$\text{Sensitivity } = \frac{TP}{TP + FN} \tag{13}$$

$$\text{Specificity } = \frac{TN}{TN + FP} \tag{14}$$

where TP, FP, TN, FN, t and f refer to true positive, false positive, true negative, false negative, true positive rate and false positive rate respectively. The AUC value represents the probability that the predicted positive samples ranks ahead of the negative samples. Because of the comprehensively considers the sensitivity and specificity of a classifier, the ISIC skin lesion classification challenge used the AUC value as a gold metrics. Accuracy is the most common evaluation metric, which measures the overall classification accuracy of positive and negative samples. Sensitivity indicates the ability to predict positive samples. In contrast, Specificity measure the ability to predict negative samples.

3.5. Main Results

We compare our framework with baseline, and report the results in Table 1. Compare to the baseline, f_s of SISLD improves the AUC, Sensitivity, Specificity, Accuracy by 1.38%, 6.72%, 0.95%, 0.76% respectively. The experimental results show that our pseudo-labeling strategy could ensure the quality of pseudo-labels and alleviate the influence of class imbalance. Additionally, SISLD model boost the performance, resulting in 0.14%, 1.83%, 0.21% absolute AUC, Sensitivity, Accuracy improvement, but the Specificity score was comparatively lower.

Table 1: Classification metrics on ISIC2018 under 20% labeled data setting.

Methods	Percentage		Evaluation Metrics			
	Labeled	Unlabeled	AUC[%]	Sensitivity[%]	Specificity[%]	Accuracy[%]
Baseline	20%	80%	92.60	67.38	91.92	92.65
f_s of SISLD	20%	80%	93.98	74.10	**92.87**	93.41
SISLD	20%	80%	**94.12**	**75.93**	92.87	**93.62**

3.6. Comparison with Benchmarks

To further evaluate the performance of our method, we compare them with several existing deep learning frameworks. Table 2 lists the lesion classification results of different frameworks, which includes SS-DCGAN(Diaz-Pinto et al., 2019), TCSE(Li et al., 2018), TE(Laine and Aila, 2016) and MT(Tarvainen and Valpola, 2017).SS-DCGAN utilize GAN to improve the network training in semi-supervised learning. TCSE, TE and MT all employ consistency regularization optimize the model. We use the supervised method as the upper bound performance and the recurrent SRC-MT trained with 20% labeled data as the baseline performance. As we can see in Table 2, SRC-MT can achieve AUC, Sensitivity, Specificity, and accuracy of 92.60%, 67.38%, 91.92% and 92.65%, both AUC and accuracy of SRC-MT are better than SS-DCGAN and TCSE. TE and MT archive higher score in AUC and Sensitivity. In contrast, our method can achieve AUC, Sensitivity, Specificity, and Accuracy of 94.12%, 75.93%, 92.87% and 93.62%, which means an increase of 1.52% of AUC, 8.55% of Sensitivity, 0.95% of Specificity and 0.97% of Accuracy. Experimental results showed that our framework could better utilize the existing labeled samples and unlabeled samples to boost the performance. The main reason is that, in the pseudo-labeling process, we set a threshold to ensure the quality of the pseudo labels, and change the distribution of pseudo labels to reduces the interference of class imbalance on the model. In the distillation process, we utilize the model and pseudo-labels from the previous process to achieve model performance improvement through self-distillation.

Table 2: Classification metrics on ISIC2018 under 20% labeled data setting.

Methods	Percentage		Evaluation Metrics			
	Labeled	Unlabeled	AUC[%]	Sensitivity[%]	Specificity[%]	Accuracy[%]
Upper Bound	100%	0%	95.43	75.20	94.94	95.10
Baseline (SRC-MT)	20%	80%	92.60	67.38	91.92	92.65
SS-DCGAN	20%	80%	91.28	67.72	92.56	92.27
TCSE	20%	80%	92.24	68.17	92.51	92.35
TE	20%	80%	92.70	69.81	92.55	92.26
MT	20%	80%	92.96	69.75	92.20	92.48
SISLD	20%	80%	**94.12**	**75.93**	**92.87**	**93.62**

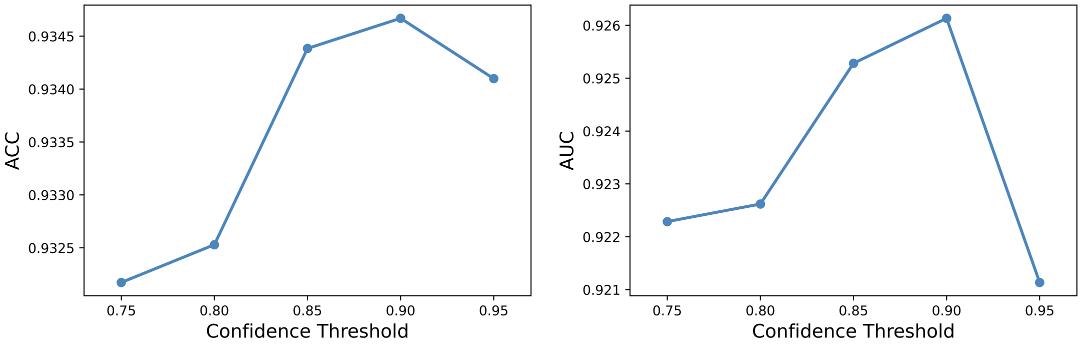
3.7. Ablation study

To evaluate and understand the contribution of each critical component in our method, we implement ablation experiments. The experiments in this section are all performed with 20% labeled data.

Effect of confidence threshold τ. To investigate the impact of the confidence threshold τ, we implemented ablation studies under different scores on the validation set, and let P_n=700. The corresponding results are shown in Table 3. As illustrated in Fig. 4, as the increases of τ , the AUC steadily rises, from 92.22% to 92.61%. A similar observation can be also found on Accuracy score, and increases in pace with the growth of and peaks at 0.90. However, keep enlarging the value of τ will lead to a performance degradation, when exceeds 0.90, the Accuracy and AUC decrease significantly. When introducing high confidence pseudo-labels of majority class, incorrect pseudo-labels amplify the misperceptions of model. As a result, we fix the value of τ as 0.90 in the following experiments.

Table 3: Results of different confidence threshold τ on the ISIC2018.

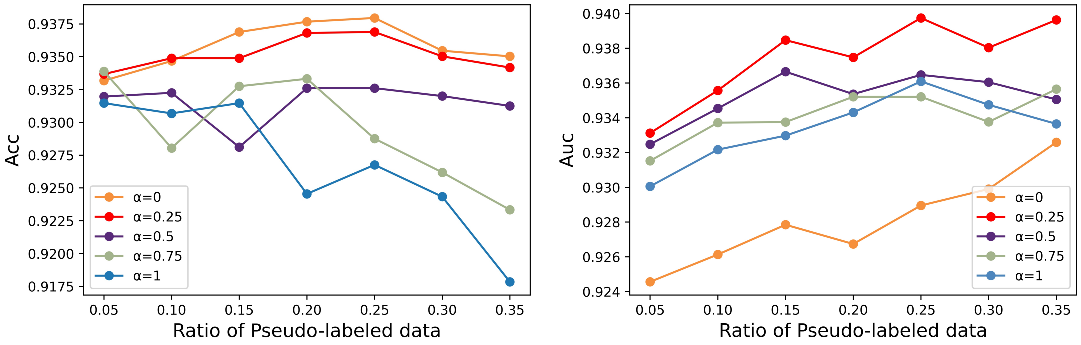
Threshold	Percentage		Evaluation Metrics	
	Labeled	Unlabeled	AUC[%]	Accuracy[%]
0.75	20%	80%	92.22	93.22
0.80	20%	80%	92.26	93.25
0.85	20%	80%	92.52	93.44
0.90	20%	80%	**92.61**	**93.47**
0.95	20%	80%	92.11	93.40



(a) (b)

Figure 4: The influence of confidence threshold τ on accuracy and AUC.

Effect of balanced factor α. SISLD introduce a balanced factor α that control the distribution of selected pseudo-labels. In Fig. 5, we show how α influence the performance over generations. In this section, the experiments are all performed with SISLD on ISIC2018 under 20% (1400) labeled data setting. When α= 0, the method falls back to conventional pseudo-labeling with the distribution of pseudo-labels equals to labeled data. As we can see, conventional pseudo-labeling strategies have yielded the best results on accuracy. This is since our strategy forces the model to improving the performance of minority classes. As the increases of and the number of pseudo-labels, accuracy becomes worse. The main reason for this phenomenon is that the dataset is imbalanced, and the performance gains on minority classes are much less than the performance drops on majority classes. In the other hand, the AUC of conventional pseudo-labeling strategy is always lower than class balanced strategy. In contrast, when α=0.25, our class balanced strategy can improved the AUC and maintain Accuracy stability.



(a) (b)

Figure 5: The influence of balanced factor α on accuracy and AUC.

Effect of loss function. To achieve knowledge distillation, we probe the effect of loss functions by comparing performance of only self-distillation $\mathcal{L}_{sd}$, only relation distillation $\mathcal{L}_{rd}$ and all loss functions. In Table 4, the results show that both $\mathcal{L}_{sd}$ and $\mathcal{L}_{rd}$ bring moderate improvement upon the f_s of SISLD. Our results demonstrated that use only $\mathcal{L}_{sd}$ and only $\mathcal{L}_{sd}$ can enhance the performance of baseline model in AUC and Accuracy, while $\mathcal{L}_{sd}$ achieve best performance. However, distillation with both losses did not achieve further significant improvements. We speculate that the effect of class imbalance was not eliminated during the self-distillation.

Table 4: Results of different loss function in self-distillation.

Methods	Percentage		Evaluation Metrics			
	Labeled	Unlabeled	AUC[%]	Sensitivity[%]	Specificity[%]	Accuracy[%]
f_s of SISLD	20%	80%	93.98	74.10	92.94	93.41
w/$\mathcal{L}_{sd}$	20%	80%	**94.12**	75.93	92.87	**93.62**
w/$\mathcal{L}_{rd}$	20%	80%	94.08	75.40	92.27	93.44
w/$\mathcal{L}_{sd} + \mathcal{L}_{rd}$	20%	80%	94.01	**76.21**	**93.03**	93.38

4. Conclusion

In this work, we propose a novel self-improving framework for skin lesion diagnosis. The proposed framework includes two process. First, we joint consistency regularization and pseudo-labeling, and achieve superior performance to existing SSL methods. Second, we utilize the optimal model of pseudo-labeling process to guide the student model of self-distillation process. By combining two self-improving methods on consistency regularization framework, which are pseudo-labeling and self-distillation, our framework improve the performance of skin lesion diagnosis without additional labels. The overall results on the ISIC2018 dataset showed that the proposed framework could effectively utilize unlabeled data and achieve high performance.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (NSFC) (Grant No.:61861004, 61862006).

References

Jack Burdick, Oge Marques, Janet Weinthal, and Borko Furht. Rethinking skin lesion segmentation in a convolutional classifier. *Journal of digital imaging*, 31(4):435–440, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Andres Diaz-Pinto, Adrián Colomer, Valery Naranjo, Sandra Morales, Yanwu Xu, and Alejandro F Frangi. Retinal image synthesis and semi-supervised learning for glaucoma assessment. *IEEE transactions on medical imaging*, 38(9):2211–2218, 2019.

Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

SM Goldsmith and AR Solomon. A series of melanomas smaller than 4 mm and implications for the abcde rule. *Journal of the European Academy of Dermatology and Venereology*, 21(7):929–934, 2007.

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

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

Sara Hosseinzadeh Kassani and Peyman Hosseinzadeh Kassani. A comparative study of deep learning architectures on melanoma detection. *Tissue and Cell*, 58:76–83, 2019.

Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2):538–546, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, and Pheng-Ann Heng. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. *arXiv preprint arXiv:1808.03887*, 2018.

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.

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. *arXiv preprint arXiv:2101.06329*, 2021.

Adam I Rubin, Elbert H Chen, and Désirée Ratner. Basal-cell carcinoma. *New England Journal of Medicine*, 353(21):2262–2269, 2005.

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. *Advances in neural information processing systems*, 33:596–608, 2020.

Alina Sultana, Mihai Ciuc, Tiberiu Radulescu, Liu Wanyu, and Diana Petrache. Preliminary work on dermatoscopic lesion segmentation. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 2273–2277. IEEE, 2012.

Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Xiaohong Wang, Weimin Huang, Zhongkang Lu, and Su Huang. Multi-level attentive skin lesion learning for melanoma classification. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3924–3927. IEEE, 2021.

Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10857–10866, 2021.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019.