Is user feedback always informative? Retrieval Latent Defending for Semi-Supervised Domain Adaptation without Source Data

Junha Song[1,2], Tae Soo Kim[2], Junha Kim[2], Gunhee Nam[2], Thijs Kooi[2], and Jaegul Choo[1]

[1] KAIST {sb020518, jchoo}@kaist.ac.kr
[2] Lunit Inc. {junha.kim, taesoo.kim, ghnam, tkooi}@lunit.io

Abstract. This paper aims to adapt the source model to the target environment, leveraging small user feedback (*i.e.*, labeled target data) readily available in real-world applications. We find that existing semi-supervised domain adaptation (SemiSDA) methods often suffer from poorly improved adaptation performance when directly utilizing such feedback data, as shown in Figure 1. We analyze this phenomenon via a novel concept called *Negatively Biased Feedback* (NBF), which stems from the observation that user feedback is more likely for data points where the model produces incorrect predictions. To leverage this feedback while avoiding the issue, we propose a scalable adapting approach, *Retrieval Latent Defending*. This approach helps existing SemiSDA methods to adapt the model with a balanced supervised signal by utilizing latent defending samples throughout the adaptation process. We demonstrate the problem caused by NBF and the efficacy of our approach across various benchmarks, including image classification, semantic segmentation, and a real-world medical imaging application. Our extensive experiments reveal that integrating our approach with multiple state-of-the-art SemiSDA methods leads to significant performance improvements.

Keywords: Rethinking user-provided feedback · Semi-supervised & Source-free domain adaptation · Medical image diagnosis

1 Introduction

While deep neural networks have demonstrated remarkable performance in the development domain (*i.e.*, source domain) [15,23], they often suffer from performance degradation in the deployed domain (*i.e.*, target domain) due to domain shift [17, 72, 78]. To mitigate this issue, domain adaptation (DA) techniques have been introduced [34, 58, 70]. The most common DA tasks include semi-supervised domain adaptation (SemiSDA) and source-free domain adaptation (SFDA). SemiSDA aims to adapt the model given a small amount of labeled target data along with massive unlabeled target data [6, 58, 66, 99]. SFDA conducts adaptation with only target data without accessing source data considering data privacy or memory constraints in edge devices [34, 67, 92].

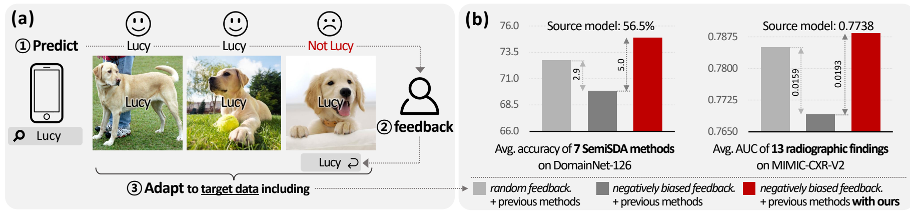


Fig. 1: (a) User feedback. Users can provide feedback while interacting with an ML product, where feedback is likely to be biased towards misclassified samples, which we define as *Negatively Biased Feedback* (*NBF*). **(b) Adaptation results.** We adapt the source model with small user feedback and large unlabeled target data using previous semi-supervised domain adaptation (SemiSDA) algorithms. Compared to *random feedback*, which is the classical SemiSDA setup where labeled data is a random subset of target data, model adaptation with NBF leads to subpar performance. This paper analyzes this problem and introduces a scalable solution.

Despite such advances in DA, adapting the model with *user feedback* still remains an open area for further research, even though practical machine learning (ML) products often allow users to provide feedback in order to further improve the model in the target environment. For example, facial recognition or medical image diagnosis applications enable users to give feedback correcting *wrong* model predictions, as depicted in Figure 1 (a). Since feedback can be modeled in this case as a small amount of labeled target data, it is anticipated that previous SemiSDA methods assuming the same setup would yield promising results. However, we observe that they show inferior adaptation performance on multiple DA benchmarks when using such user feedback in practice, as shown in the dark-gray bar ■ in Figure 1 (b).

We introduce a novel concept called *Negatively Biased Feedback* (NBF) to explain this phenomenon. NBF is based on the observation that user feedback is more likely to be derived from *incorrect* model predictions. For example, a radiologist might log a misdiagnosed chest X-ray by the model, as its accuracy directly impacts the patient's survival. Interestingly, our observation aligns with findings from cognitive psychology literature [3,57] that proves that humans are more likely to react and provide feedback to negative events (*i.e.*, wrong model predictions). Since such an NBF scenario is feasible, we analyze its unexpected impact on SemiSDA observed above. We identify that a biased distribution of NBF within the overall data distribution leads to sub-optimal adaptation results, particularly compared to Random Feedback (RF). RF represents the classical SemiSDA setup, where labeled data is randomly selected from the target data.

To address the problem caused by NBF, we present a *scalable* approach named *Retrieval Latent Defending*, which can be seamlessly integrated with existing SemiSDA methods. Our approach allows them to adapt the model without a strong dependence on the biasedly distributed labeled data. Specifically, we balance the supervised adapting signal by appending latent defending samples to the mini-batch and help to keep the model's balanced class discriminability throughout adapting iterations. We evaluate the unexpected influence of NBF using various benchmarks, including image classification, semantic segmentation,

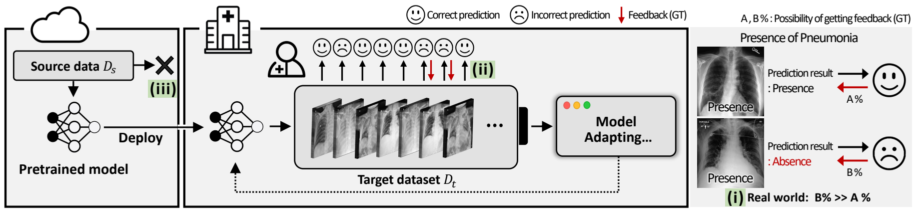


Fig. 2: Adaptation with user feedback can be effective in alleviating performance degradation caused by domain shift. However, there are some challenges: (i) user feedback may be a biased sampling of the true target distribution due to the nature of feedback, (ii) the amount of the ground truths (GT) labels obtained through feedback is small, and (iii) only unlabeled target data is typically available, not source data.

and medical image diagnosis. Building upon these evaluations, we demonstrate that our approach not only complements, but significantly enhances the performance of multiple SemiSDA methods.

The contributions of the paper are as follows:

- We introduce the novel concept called *Negatively Biased Feedback* and uncover that it can lead to sub-optimal adaptation performance of existing SemiSDA methods.
- We analyze this problem and present a scalable solution, *Retrieval Latent Defending*, that combines with SemiSDA methods and allows them to avoid the unexpected effect of NBF.
- We show that our approach generalizes through diverse DA benchmarks and improves adaptation results of state-of-the-art SemiSDA methods.
- We publicly release the code on `https://github.com/junha1125/RLD-SemiSDA`.

2 Related Work

Adaptation in the deployment environment. Real-world ML products often encounter performance degradation caused by gaps between the source and target environment [17]. One solution is to adapt the model using unlabeled data observed in the target domain, referred to as **unsupervised domain adaptation** (UDA) [37, 59, 72]. Works on UDA use both source and target data to improve the target performance by using methods such as domain discrepancy minimization by adversarial training [18, 41, 59, 70, 72, 73, 76], and self-training with pseudo labels [45, 51, 97, 98]. **Source-free DA** (SFDA) builds on UDA and imposes an additional constraint that the source data can not be accessed during domain adaptation. This has practical implications for addressing data privacy concerns or barriers in data transmission to edge devices [34, 38, 77, 95]. The majority of recent SFDA works rely on strategies like domain clustering [34], nearest neighbors [91–93], and contrastive learning [8, 35, 101]. Nevertheless, SFDA does not consider the availability of small labeled data, which may be available in practical ML systems. **Semi-supervised DA** (SemiSDA) works mainly demonstrate that permitting small labeled data in the target domain can substantially enhance adaptation performance compared to traditional UDA [58].

Their primary strategy is to use domain alignment [20, 33, 58, 94], multi-view consistency [2, 6, 33, 89], and asymmetric co-training [36, 90].

Active domain adaptation (ActiveDA) [25, 55, 86] envisions a scenario in which the machine selects specific target samples and instructs annotators to label them. The primary objective of ActiveDA is to strategically identify and select the most informative samples for annotation. These chosen samples (*i.e.*, labeled target data) are subsequently utilized to update the source model using SemiSDA methods [33, 58], and the effectiveness of ActiveDA is assessed by evaluating the target performance of the adapted model.

Semi-supervised learning (SemiSL) aims to reduce expensive human annotations, and propose methods to train a model from scratch using massive unlabeled data along with limited amounts of labeled data [43, 74]. The majority of SemiSL methods depend on consistency regularization [4, 5, 16, 60, 66, 87], which helps the model to make similar predictions for augmented versions of the same image. Moreover, adaptive thresholding [9, 12, 24, 66, 81, 88, 99] is also popularly utilized to produce reliable pseudo labels from unlabeled data.

SemiSDA and SemiSL setups mimic small labeled datasets by randomly selecting subsets of the target dataset, whereas ActiveDA involves selections instructed by the machine. In contrast, this paper posits that in real-world applications, labeled data is typically acquired through user intervention. Additionally, users often provide feedback on samples misclassified by the model (*i.e.*, negatively biased feedback), a process detailed in the following section.

	UDA	SFDA	ActiveDA	SemiSDA	SemiSL	Our setup
Adaptation	∘	∘	∘	∘	×	∘
Source-free	×	∘	×	×	-	∘
Feedback	×	×	machine-instructed	randomly selected	randomly selected	**user-provided**

The table above summarizes the comparison of relevant studies to our setup. In the table, adaptation means fine-tuning the source pre-trained model (as opposed to training from scratch); feedback represents a small number of labeled target samples. Appendix A provides further comparisons with settings like class-imbalanced SemiSDA and test-time adaptation (TTA).

3 Negatively Biased Feedback

3.1 Adaptation with user feedback.

Our adaptation setup is illustrated in Figure 2. A model is pre-trained on the source data D_s. Next, the model is deployed to the target domain, such as a smartphone or a hospital, where we assume the transfer of D_s is prohibited due to data privacy regulations or resource constraints (same setup as SFDA [34]). While users utilize ML products on the target domain, the model provides prediction results for data observed in the target domain D_t and occasionally obtains user feedback in the form of annotations y. We represent the target data as $D_t = X_t^{lb} \cup X_t^{ulb}$, where $X_t^{lb} = \{(x_{lb}^n, y_{lb}^n) : n \in [1..N_{lb}]\}$, $X_t^{ulb} = \{(x_{ulb}^n) : n \in [1..N_{ulb}]\}$, x_{lb} and x_{ulb} denote labeled and unlabeled data and N_{lb} and N_{ulb} is their number of data. Lastly, the model can utilize D_t and SemiSDA algorithms for adaptation during its inactive phase (*e.g.*, when users do not use the product,

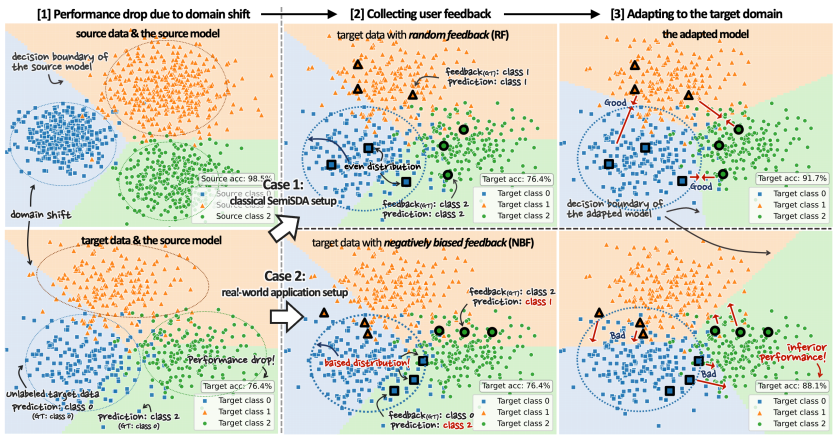


Fig. 3: Effect of negatively biased feedback. Our novel observations are that ***(a)*** user-provided feedback in practice has a biased distribution in each class cluster (the bottom center sub-figure) which is in contrast to random feedback, ***(b)*** Existing SemiSDA methods adapt the model by dominating the labeled data points (the right sub-figures) even though they are biasedly positioned, and ***(c)*** NBF prevents the model from having a decision boundary for true class clusters and leads to inferior adaptation performance (the bottom right sub-figure).

like at nighttime) in order to alleviate performance degradation due to domain shift or to personalize the model based on user feedback.

Rethinking user-provided feedback. Classical SemiSDA works simply assume that a random subset in target data D_t is labeled by users when building X_t^{lb}. However, as illustrated in Figure 2 (i), we suggest that users are more likely to provide feedback on misclassified samples by the source model, named negatively biased feedback (NBF). This behavior can be understood from two perspectives: (a) users generally expect their feedback to be used as a basis of model improvement, motivating them to provide NBF, and (b) humans tend to react more strongly to negative experiences, such as receiving incorrect predictions, as observed in psychological studies [3,57]. We note that the NBF assumption holds more strongly for the *medical* application: it is reasonable to imagine that the user (*i.e.*, radiologist) logs the *mistakes* of the model while diagnosing a chest X-ray exam because the diagnostic accuracy of the model is directly related to the patient's chances of survival. Furthermore, applications beyond the medical domain can also exhibit NBF. For instance, users in self-driving cars can report errors, such as object detection failures or navigation mistakes, to enhance the car's driving capabilities.

3.2 Influence of NBF on SemiSDA

Simulation study. As shown in Figure 3, we conduct a simulation study to understand the effect of NBF on SemiSDA. We first use the blobs dataset [53] and construct the source and target data so that domain shift exists between them (left sub-figures). We pre-train a source model on the source data and compute the accuracy in the target domain, where the performance drop due to domain shift is observed (98.5%→76.4%). Next, we simulate two types of feedback (*i.e.*, labeled data): random feedback and negatively biased feedback following a previous SemiSDA setup and our setup, respectively. Specifically, NBF is randomly selected among misclassified samples by the source model. We find that random feedback (RF) points are *evenly* distributed, while NBF points are *biasedly* positioned across each class cluster (refer to blue points in the dashed circle in the center sub-figures).

To alleviate the performance drop caused by domain shift, we adapt the model using the target data and a semi-supervised method, Pseudo-labeling [1]. This method iteratively optimizes the model by the cross-entropy loss computed by the ground truth of labeled data and pseudo labels of unlabeled data in a mini-batch (pseudo labels are predicted by the *current* adapting model so they can be changed according to an updating decision boundary. Further comprehension can be achieved by referring Appendix B.). The SemiSDA results are shown in the right sub-figures, where we make two interesting observations: (i) the distribution of labeled data can contribute significantly to a decision boundary of the adapted model (red arrows in the figure), and (ii) the adapted model under NBF has poorly improved performance compared with one under RF (76.4%→88.1% with NBF, but 76.4%→91.7% with RF).

Unexpected influence of NBF. Our intuitive reasoning probably suggests that NBF provides more information than RF by correcting more source model deficiencies, and thus leads to better adaptation performance. However, we empirically show that NBF can result in inferior adaptation performance due to its biased distribution across each class cluster, as illustrated in Figure 3. Surprisingly, we also show that this problem persists, even with other state-of-the-art SemiSDA methods and large datasets for various DA benchmarks, including image classification, semantic segmentation, and medical image diagnosis. Our work highlights the importance of careful design when using user feedback in real-world scenarios and, to the best of our knowledge, is the first study to uncover and analyze this phenomenon.

4 Approach

4.1 Prerequisite: Previous SemiSDA method

Previous SemiSDA and SemiSL works typically construct a mini-batch with labeled data $\{(x^b_{lb}, y^b_{lb}) : b \in [1..B]\}$, and unlabeled data whose size is μ times larger than labeled ones $\{(x^b_{ulb}) : b \in [1..\mu{\cdot}B]\}$, where B is the mini-batch size for labeled data. To adapt the model iteratively, they compute the cross-entropy loss $\mathcal{H}(\cdot,\cdot)$ with labeled data and the consistency regularization to multi-view of unlabeled data, which are formulated as the following:

$$\mathcal{L}_{sup} = \frac{1}{B}\sum_{b=1}^{B}\mathcal{H}(y^b_{lb}, f_\theta(x^b_{lb})), \quad \mathcal{L}_{unsup} = \frac{1}{\mu \cdot B}\sum_{b=1}^{\mu \cdot B}\mathcal{H}(\hat{y}^b_{ulb}, f_\theta(\Omega(x^b_{ulb}))), \tag{1}$$

where $f_\theta(\cdot)$ is the output probability from the model, $\hat{y}_{ulb}$ denotes a pseudo label obtained from $f_\theta(\omega(x_{ulb}))$, and $\omega(\cdot)$ and $\Omega(\cdot)$ represent weak and strong image

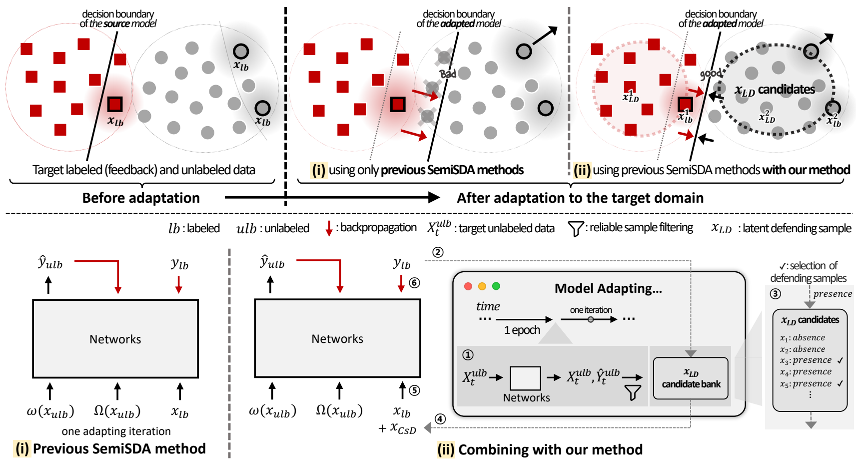


Fig. 4: Even though labeled data (x_{lb}, y_{lb}) is biasedly positioned, the model needs to be adapted with balanced class discriminability (*i.e.*, decision boundary). **(i)** However, previous SemiSDA methods have overlooked this fact and used the labeled data naively by applying a cross-entropy loss, leading to inadequate adaptation performance. **(ii)** To alleviate this problem, we propose a scalable adapting approach, retrieval latent defending, which allows the model to adjust the balance of a mini-batch on each iteration by using latent defending samples x_{LD} together with labeled data x_{lb}.

augmentation, respectively. While sharing the core framework, each SemiSDA method employs distinct adapting strategies, especially to enhance the effectiveness of the use of *unlabeled* data rather than *labeled* data [6, 81, 96].

Problem of previous works. Since previous SemiSDA methods have overlooked the unexpected impact of NBF, they often suffer from sub-optimal performance under the NBF assumption (shown in Section 5). To address this problem, we focus on developing a *scalable* solution that (i) can easily combine with existing DA methods without modifying their core adapting strategies and (ii) can be applied to a wide range of benchmarks, including medical image diagnosis.

4.2 Retrieval Latent Defending

Based on the observations in Figure 3, we illustrate the unintended effect of NBF when using an existing SemiSDA method in Figure 4 (top center), where NBF is likely to exhibit a biased distribution, leading to undesirable adaptation results. To alleviate this issue, we propose *Retrieval Latent Defending* as depicted in Figure 4 (bottom). ① Prior to each epoch, we generate a candidate bank of data points, denoted as x_{LD}. ②∼④ For each adapting iteration, we balance the mini-batch by retrieving latent defending samples x_{LD} from the bank. ⑤∼⑥ The model is then adapted using the reconfigured mini-batch and following the baseline SemiSDA approach. We hypothesize that the latent space progressively created by the x_{LD} candidates throughout the adaptation process (bold dashed circle in Figure 4 (top right)) mitigates the issue caused by NBF, thereby allowing the SemiSDA baseline to achieve robust adaptation against NBF.

Candidate bank generation. The candidate bank serves as a repository of pseudo labels $\hat{Y}_t^{ulb}$ for a subset of the target unlabeled data X_t^{ulb}. Before each epoch, we freeze the model and use it to generate pseudo labels $\hat{Y}_t^{ulb}=\{(\hat{y}_{ulb}^n) : n \in [1..N_{ulb}]\}$, where $\hat{y}_{ulb}^n$ is assigned to x_{ulb}^n as the predicted class with the highest softmax probability: $\hat{y}_{ulb}^n = \text{argmax}_c \left[f_\theta(x_{ulb}^n)\right]_c$. We then retain only samples with the top $p\%$ highest probabilities within each class. This filtering step helps mitigate the inclusion of data with potentially inaccurate pseudo labels, as the model's predictions on X_t^{ulb} might not always be perfect.

Defending sample selection. We select k latent defending samples x_{LD} from the bank at random for each labeled data (x_{lb}^b, y_{lb}^b). These selected samples share the *same* pseudo label as the ground-truth label of their associated counterparts (*i.e.*, $\hat{y}_{LD} = y_{lb}^b$). By incorporating these defending samples, we balance the data distribution within the current mini-batch. For example, consider x_{lb}^1 and x_{lb}^2 in Figure 4 (top right). As these labeled samples are included in the current mini-batch alongside the selected defending samples x_{LD}^1 and x_{LD}^2, we expect to prevent the supervised adapting signal from becoming overly dependent on the labeled samples. We imagine the effect of the defending samples *throughout* the adaptation process and depict the latent space formed gradually by the x_{LD} candidates as bold dashed circles in Figure 4 (top right).

Consequently, the overall loss consists of the sum of losses in Eq. (1) and a loss from our proposed method as,

$$\mathcal{L}_{total} = \underbrace{\mathcal{L}_{sup} + \mathcal{L}_{unsup}}_{\text{baseline}} + \underbrace{\frac{1}{k \cdot B} \sum_{b=1}^{k \cdot B} \mathcal{H}(\hat{y}_{LD}^b, f_\theta(x_{LD}^b))}_{\text{retrieval latent defending}} . \tag{2}$$

Importance of our method. Understanding the impact of NBF on adaptation performance is crucial. For example, naively adapting a model for a medical application using radiologist-provided feedback can actually cause performance degradation (shown in Table 5), potentially posing significant risks to patients. We propose a *scalable* and *simple* approach to solve the problem caused by NBF, which can not be addressed by existing methods. Given the practicality of the NBF problem and the scalability of our solution, we believe our work holds considerable potential for real-world applications.

5 Experiments

5.1 Experimental Setups

Our approach is simple enough to seamlessly combine with existing SemiSDA algorithms and also be applied to diverse benchmarks. This section describes our experimental setup for natural image classification tasks and a real-world medical application. Details for semantic segmentation experiments are in Appendix D.

Baselines. We validate our approach by combining various state-of-the-art algorithms for SemiSDA [58] (*e.g.*, CDAC [33] and AdaMatch [6]) and SemiSL [66,87] (*e.g.*, FlexMatch [96] and FreeMatch [81]). Note that the SemiSL methods have been demonstrated to be strong *SemiSDA learners* [99], so we can consider them

	method	feedback	average	r→c	r→p	p→c	c→s	s→p	r→s	p→r
ResNet	AdaMatch [6]	RF	67.6	66.6	68.5	68.5	60.3	69.2	58.7	81.5
		NBF	64.5 (-3.1)	64.3	66.1	65.6	56.9	65.6	54.2	78.9
	w/ ours	NBF	**72.0** (+7.5)	74.5	72.7	73.9	65.5	70.0	64.3	83.2
ViT	AdaMatch [6]	RF	74.7	75.3	76.9	73.8	68.0	76.3	67.1	85.5
		NBF	73.7 (-1.0)	74.7	76.2	74.7	65.7	74.0	66.8	84.0
	w/ ours	NBF	**75.9** (+2.2)	76.9	77.8	77.8	68.5	76.6	68.3	85.1

Table 1: Adaptation results on DomainNet-126. We simulate seven domain-shift scenarios (*i.e.*, source → target). The model is pre-trained on the source domain and then adapted to a training set of the target domain. The results on the test set of the target domain are reported as the top-1 accuracy (%). DomainNet-126 [54,58] dataset includes real, painting, sketch, and clip-art domains. In this experiment, we assume that the 378 feedback samples (*i.e.*, 3 labeled data per class) are obtained from users. A state-of-the-art SemiSDA method, AdaMatch [6], is used as a baseline.

method	feedback	average	a → c	a → p	a → r	c → a	c → p	c → r	p → a	p → c	p → r	r → a	r → c	r → p
AdaMatch [6]	RF	70.9	55.4	80.4	75.9	65.7	81.5	74.6	65.9	58.7	78.4	68.8	61.5	84.3
	NBF	69.3 (-1.6)	54.2	76.6	75.3	65.9	79.3	75.5	63.7	57.4	75.9	66.7	56.8	84.2
w/ ours	NBF	**73.8** (+4.5)	62.2	81.0	79.7	68.8	85.4	78.6	67.7	61.7	79.5	69.0	64.1	88.2

Table 2: Adaptation results on OfficeHome. OfficeHome [75] dataset includes real, product, art, and clip-art domain. We assume that the 195 feedback samples (*i.e.*, 3 labeled data per class) are obtained. AdaMatch [6] and ResNet-50 [23] are used.

as SemiSDA methods. For medical experiments, we use Pseudo-labeling [1] as a baseline since it is easily applicable to medical image adaptation.

Datasets. We utilize natural image datasets containing multiple kinds of domains (*e.g.*, real and painting). The datasets include DomainNet-126 [54,58] with 142k images of 126 classes, and OfficeHome [75] with 15K images of 65 classes.

To conduct medical experiments, we present a practical medical setting. We adopt the MIMIC-CXR-V2 dataset [27]. It assumes a multi-finding binary classification setup, where multiple radiographic findings, like Pneumonia and Atelectasis, can coexist in a single chest X-ray (CXR) sample. Thus, the model predicts the presence or absence (binary classes) of *each individual* finding. We simulate domain shift by using Posterior-Anterior (PA)-view data as the source and AP-view data as the target, capturing real-world variations in data acquisition. Typically, patients requiring an AP X-ray are those facing positioning challenges that prevent them from undergoing a PA X-ray. Therefore, this setup can be seen as a scenario where the target environment is the intensive care unit, which hospitalizes critically ill patients.

Following the recent SemiSDA [94] and SFDA [8] setups, we assume the model is pre-trained in the source domain and deployed in the target domain. Since the datasets above were not initially divided into training and test sets, we performed a random 8:2 split within each domain, designating them respectively for training and testing. The training set is used to adapt the model, while the test set is used to report the top-1 accuracy.

User feedback. Feedback given by users is modeled as annotations $\{y_{lb}^n : n \in [1.. N_{lb}]\}$ on a small subset of the target's training set D_t^{train}, while the remaining of them are used as unlabeled target data. In our experiments, we take into account two types of feedback: random feedback (RF) and negatively biased feedback (NBF). RF is the same setup of classical SemiSDA and SemiSL, where randomly selected samples from D_t^{train} are used as small labeled set X_t^{lb}. For NBF, we randomly select samples that are incorrectly predicted in D_t^{train} by

	feed. amount	378 (3 labeled data per class)			630 (5 labeled data per class)		
	method	RF	NBF	w/ ours	RF	NBF	w/ ours
	Source model	56.5					
ResNet-50 [23]	MME [58]	69.5	68.4 (-1.1)	70.8 (+2.4)	71.2	70.1 (-1.1)	72.5 (+2.4)
ResNet-50 [23]	CDAC [33]	68.3	64.6 (-3.7)	73.2 (+8.6)	71.7	68.1 (-3.6)	74.9 (+6.8)
ResNet-50 [23]	AdaMatch [6]	67.6	64.5 (-3.1)	72.0 (+7.5)	70.9	67.7 (-3.2)	74.3 (+6.6)
ResNet-50 [23]	FixMatch [66]	67.6	63.4 (-4.2)	73.2 (+9.8)	71.5	66.1 (-5.4)	75.1 (+9.0)
ResNet-50 [23]	UDA [87]	69.2	64.9 (-4.3)	73.4 (+8.5)	72.9	68.8 (-4.1)	75.3 (+6.5)
ResNet-50 [23]	FlexMatch [96]	73.3	71.4 (-1.9)	74.7 (+3.3)	75.3	73.9 (-1.4)	76.0 (+2.1)
ResNet-50 [23]	FreeMatch [81]	73.8	72.0 (-1.8)	**74.8** (+2.8)	75.6	74.4 (-1.2)	**76.1** (+1.7)
	Fully supervised	83.6					
	Source model	64.5					
ViT-S [15]	MME [58]	73.2	72.7 (-0.5)	74.1 (+1.4)	74.5	74.0 (-0.5)	75.2 (+1.2)
ViT-S [15]	CDAC [33]	74.2	72.8 (-1.4)	75.4 (+2.6)	75.4	74.1 (-1.3)	76.2 (+2.1)
ViT-S [15]	AdaMatch [6]	74.7	73.7 (-1.0)	**75.9** (+2.2)	75.9	75.1 (-0.8)	76.7 (+1.6)
ViT-S [15]	FixMatch [66]	74.6	73.0 (-1.6)	75.6 (+2.6)	75.7	74.3 (-1.4)	76.5 (+2.2)
ViT-S [15]	UDA [87]	74.8	73.3 (-1.5)	75.8 (+2.5)	75.9	74.5 (-1.4)	76.7 (+2.2)
ViT-S [15]	FlexMatch [96]	74.9	73.9 (-1.0)	75.8 (+1.9)	76.0	75.1 (-0.9)	**76.9** (+1.8)
ViT-S [15]	FreeMatch [81]	74.9	73.9 (-1.0)	75.7 (+1.8)	76.0	75.1 (-0.9)	76.8 (+1.7)
	Fully supervised	85.4					

Table 3: Comparisons on DomainNet-126. We evaluate our method by integrating it with *SemiSDA* and *SemiSL* methods. The average accuracy of seven domain-shift scenarios in Table 1 is reported. Source model represents the pre-trained model without adaptation. Fully supervised means the model is adapted with fully labeled target data.

feed. amount	195 (3 labeled data per class)			325 (5 labeled data per class)		
method	RF	NBF	w/ ours	RF	NBF	w/ ours
Source model	57.6					
MME [58]	71.2	70.2 (-1.0)	73.4 (+3.2)	73.5	73.1 (-0.4)	75.6 (+2.5)
CDAC [33]	71.2	69.0 (-2.2)	74.3 (+5.3)	73.5	72.3 (-1.2)	75.7 (+3.4)
AdaMatch [6]	70.9	69.3 (-1.6)	73.8 (+4.5)	73.4	72.7 (-0.7)	75.5 (+2.8)
FixMatch [66]	71.4	68.6 (-2.8)	73.7 (+5.1)	73.9	72.2 (-1.7)	75.3 (+3.1)
UDA [87]	72.2	69.5 (-2.7)	74.1 (+4.6)	74.4	73.0 (-1.4)	76.0 (+3.0)
FlexMatch [96]	73.7	72.1 (-1.6)	74.7 (+2.6)	75.9	74.9 (-1.0)	**76.6** (+1.7)
FreeMatch [81]	74.0	72.7 (-1.3)	**74.8** (+2.1)	75.8	75.0 (-0.8)	**76.6** (+1.6)
Fully supervised	87.4					

Table 4: Comparisons on OfficeHome. The average accuracy of twelve domain-shift scenarios in Table 2 is reported. ResNet-50 is used.

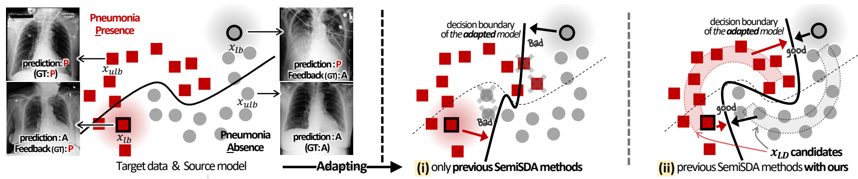
the source model (*i.e.*, the pre-trained model before adaptation). Note that we focus on the impact of a biased label distribution within the *same* class, as shown in Figure 3, and thus take the same number of feedback for each class. Further discussion about the imbalance in the number of feedback *between classes* presented in [31, 49, 83] is provided in Appendix A.2.

Network architectures. We adopt commonly used networks, ResNet [23] and ViT [15] for natural image tasks and DenseNet [26] for a medical task. We employ ResNet-50 with the last classification layer comprising a weight normalization layer and a bottleneck layer, following previous works [8, 34] and use the ViT-Small (*i.e.*, ViT-S) introduced in [80]. The DenseNet-121 is used, provided in TorchXrayVision [13], like existing medical works [32, 44].

Implementation details. We implement our framework by extending the publicly available USB [80] repository. Both pre-training and adaptation are conducted with a mini-batch size of 128 and the SGD optimizer. Diverse baselines for SemiSDA and SemiSL are used to compute the losses in Eq. (1). The hyper-parameters for each baseline simply follow USB [80] or public code [33, 58]. For all experiments, our approach uses the same hyper-parameters of the appended defending samples k and reliable filtering rate p as 3 and 0.4, respectively.

5.2 Main Results

Natural image classification. Following recent DA works [8, 94], we conduct experiments on seven and twelve domain shift scenarios provided with the



method	feedback	average	atelectasis	cardiomegaly	consolidation	edema	enlarged cardio.	fracture	lung lesion	lung opacity	effusion	pleural	pneumonia	pneumothorax	support device
Source mo.		.7738	.7784	.7919	.8236	.8500	.7646	.6642	.7555	.7818	.8271	.8288	.7535	.6894	.7500
PseudoL [1]	RF	.7850	.7828	.7965	.8453	.8615	.7639	.6832	.7598	.7947	.8333	.8565	.7702	.6957	.7622
	NBF	.7691	.7719	.7851	.8202	.8468	.7403	.6934	.7446	.7809	.8070	.8260	.7521	.6979	.7324
	gap	-.0159	-.0109	-.0114	-.0252	-.0147	-.0236	+.0102	-.0152	-.0138	-.0262	-.0304	-.0181	+.0022	-.0298
w/ ours	NBF	**.7884**	.7895	.7956	.8515	.8606	.7730	.6821	.7599	.7973	.8445	.8611	.7753	.6851	.7736
	gain	+.0193	+.0176	+.0105	+.0313	+.0138	+.0326	-.0113	+.0153	+.0164	+.0375	+.0351	+.0232	-.0128	+.0412
	NBF-CE	.7639	.7682	.7834	.8124	.8418	.7403	.6808	.7472	.7744	.8005	.8199	.7469	.6879	.7277
	gap	-.0211	-.0146	-.0131	-.0330	-.0198	-.0236	-.0024	-.0126	-.0203	-.0328	-.0366	-.0233	-.0079	-.0344
w/ ours	NBF-CE	.7875	.7895	.7956	.8515	.8606	.7730	.6731	.7599	.7973	.8445	.8611	.7753	.6831	.7736
	gain	+.0236	+.0213	+.0122	+.0391	+.0189	+.0327	-.0077	+.0126	+.0229	+.0440	+.0412	+.0284	-.0048	+.0459
Fully super.		.8117	.8150	.8277	.8758	.8820	.7984	.6949	.7750	.8200	.8725	.8441	.8044	.7398	.8025

Table 5: Adaptation in a medical application. We use samples with PA-view as the source data and samples with AP-view as the target data in MIMIC-CXR-V2 dataset [27]. NBF-CE represents a scenario when NBF is composed of cases with confident errors. We use DensNet-121 [13, 26] and assume the 20 feedback for the absence and presence per finding.

DomainNet-126 and OfficeHome datasets, respectively. Table 1 and Table 2 show the results, where AdaMatch [6] is used as the baseline. We observe consistent results with Figure 3 *even on* large natural datasets: when simply applying the baseline under the NBF assumption, the adapted model shows inferior performance for most domain shifts than applying it under RF, *e.g.*, 64.5<67.6. Combining our approach with the baseline mitigates this issue and achieves a performance increase, *e.g.*, 64.5→72.0.

We also use other promising baselines and report the average accuracy of all domain shifts in Table 3 and Table 4 (all results can be found in Appendix F). While both feedback types bring performance improvement from the source model, lower performance is observed with NBF. Our method enables the baselines to not only address this problem but surpass performance under RF. The above results suggest that the biased distribution of labeled samples, which has been overlooked in previous SemiSDA works, is actually problematic, and our retrieval latent defending approach is effective.

Medical image diagnosis. Table 5 shows the results (bottom) and also depicts the effect of NBF (top center). We report the AUROC [7] for each finding following standard practice for measuring computer-aided-diagnosis model evaluation [32, 44]. The baseline SemiSDA method under NBF exhibits inferior performance compared to one under RF, but this issue can be mitigated by combining our approach.

In addition, we propose an interesting and practical scenario named NBF with more confident errors (**NBF-CE**). In this scenario, we assume that a radiologist is likely to give feedback when the model makes *confidently wrong* predictions. Imagine that the model predicts a 1% likelihood of cancer in a CXR image, but the person actually has cancer. Such failure to detect potential patients early on can significantly reduce the patient's chances of survival, so a radiologist may

labeling type	feed. amount	RF	NBF	NBF w/ ours	ENT [62]	ENT w/ ours
IAST [46, 63]	PA, 40 points	55.3	53.0 (-2.3)	**56.3** (+3.3)	53.5	56.0 (+2.5)
RIPU [85]	PA, 40 points	57.6	54.5 (-3.1)	**58.0** (+3.5)	54.6	57.7 (+3.1)

Table 6: Adaptation on semantic segmentation. The GTA5 [56] → Cityscapes [14] setup is used [72]. The target performance of the source model is 36.6 mIoU.

provide feedback to the model. To simulate NBF-CE, we select samples where the source model most confidently predicts a finding to be absent ($\hat{y}{\approx}0$) although it is clearly visible in the radiograph ($y{=}1$), and vice versa, *i.e.*, samples of $\hat{y}{\approx}1$ but $y{=}0$. Table 5 also shows the results under an NBF-CE scenario, where the model's adaptation performance is further reduced compared with NBF (0.7691 for NBF → 0.7639 for NBF-CE). By combining our method, we observe performance improvements for both NBF variants, *e.g.*, 0.7639 for NBF-CE → 0.7875 with ours. We illustrate the hypothesized impact of our method in Table 5.

Semantic segmentation. We evaluate the influence of NBF and our approach on a semantic segmentation task. We utilize the most common adaptation benchmark of GTA5 [56] to Cityscapes [14]. The baseline DA algorithms are used as IAST [46, 63] and RIPU [85] in a source-free scenario. We regard Pixel-based Annotation (PA) in which we assume 40 pixels per image like LabOR [63]. Table 6 shows results similar to those we observed in the classification and medical imaging tasks. The baselines under NBF exhibit inferior performance compared to those under RF (54.5 for NBF<57.6 for RF), but this issue is addressed by combining our approach with them (+3.5 mIoU). Although out of our scope (refer to Appendix A.1), we validate one active labeling strategy ENT [62], which assigns highly uncertain (*i.e.*, *probably misclassified*) pixels as feedback. Consequently, the feedback instructed by ENT is *biasedly* distributed in a manner similar to NBF. ENT also causes unexpected results (54.6 for ENT<57.6 for RF), and our approach alleviates this issue (+3.1 mIoU).

5.3 Ablation Study

Positive vs. Negative feedback. We study the role of feedback on the adaptation results by varying feedback configurations. Let positively-provided feedback (PF) be obtained from samples that the source model *correctly* predicts, as opposed to negatively-provided feedback (NF). We adjust the ratio of PF:NF while keeping the total number of labeled samples constant, as shown in Figure 5.

When using only FreeMatch (gray dot-dashed line), both biased feedback types (*i.e.*, NBF and PBF) result in worse adaptation performance compared to balanced feedback for the baseline, *e.g.*, 72.6 in 378:0 (PBF) < 73.3 in 252:126. In contrast, when our method is applied (red line), NBF yields the best performance. PBF and NBF can be respectively regarded as contributing previously *known* knowledge of the model and *new* knowledge that complements model deficiencies. Hence, it may be natural that NBF, which actually encodes the model's mistakes, contributes to favorable adaptation results.

Number of unlabeled samples in a mini-batch. Existing SemiSDA methods [6, 81] typically set the ratio μ between labeled and unlabeled samples in a

If not specified, we use ResNet-50 and report the average accuracy (%) of seven domain shift scenarios in Table 1 for ablation studies.

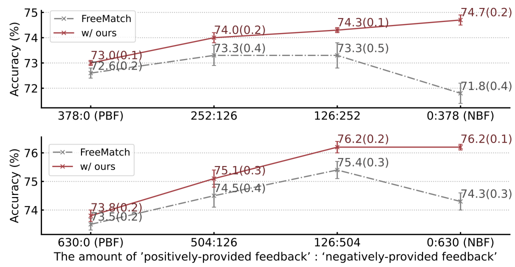


Fig. 5: NBF leads to higher performance than PBF. We compare different user-feedback configurations when the total number of feedback is 378 (top) and 630 (bottom). Positive and negative feedback refers to feedback from correct and incorrect model predictions, respectively. We run three random seed experiments and describe the average performance and standard deviation in the parenthesis.

method	feed.	negatively biased feedback (NBF)		
$\#x_{ulb}, \#x_{LD}, \#x_{lb}$		112, 0, 16	112, 48, 16	64, 48, 16
total batch size		128	176	128
FreeMatch [81]	368	72.0	74.2	**74.8** (+0.6)
AdaMatch [6]		64.5	71.3	72.0 (+0.7)
FreeMatch [81]	630	74.4	75.5	**76.1** (+0.6)
AdaMatch [6]		67.7	73.4	74.3 (+0.9)

Table 7: In the mini-batch, diminishing the number of unlabeled samples and adding our defending samples achieves better performance with our approach. We ablate them by changing the ratio μ in Section 4.1, while keeping the size of labeled samples.

mini-batch to 1:7. However, we observe that adhering to this ratio is not optimal for our approach, as shown in Table 7. Our method shows better performance when the ratio is varied to 1:4, *i.e.*, decreasing unlabeled sample sizes. This finding contradicts observations in several TTA works [28,48,67], where adaptation performance tends to increase with larger batch sizes. We speculate that it is beneficial to prioritize more reliable information, which refers to labeled data and our defending samples selected from the *filtering*-applied bank, during the adapting process. This result may be aligned with previous works for curriculum learning [40,100] and adaptive thresholding [96].

Number of labeled data. We measure the impact of feedback size (number of labeled samples) in Figure 6. The results show that the inferior performance on NBF persists even with an increased amount of feedback (gray → black line); however, our approach mitigates it and improves performance (black → red line). We make an interesting observation that the performance gap between black and red lines becomes larger as the number of available feedback decreases. Since obtaining large feedback may be challenging in real-world applications, our method is expected to be more helpful in this practical case.

Data selecting strategy. We explore various strategies for selecting defending samples to balance the mini-batch, as shown in Table 8 (top). The strategies include: in the x_{LD} candidate bank, (i) random selection regardless of the class of the labeled data, (ii) random selection in the same class as the labeled data (*i.e.*, class-aware), (iii) selecting samples close to the cluster center obtained by k means clustering [21] and (iv) selecting samples with embedded features distant from the labeled data where cosine distance is used. While our approach consistently outperforms the baseline regardless of the chosen strategy, we empirically

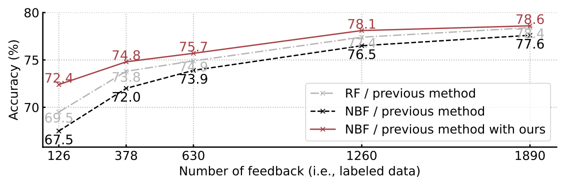


Fig. 6: More reliable adaptation with NBF. In addition to Table 3, we conduct experiments with different amounts (1,3,5,10, and 15 labeled data per class) of feedback using FreeMatch [81]. The number of available feedback is likely to be small in practice. In this case, our method achieves large performance improvement, *e.g.*, our method increases the baseline performance by +4.9 when one feedback per class is available.

selection strategy		random	random	kmeans	cosine	baseline
class-aware		✗	✓	✓	✓	-
FreeMatch [81]	Res.	74.1	**74.8**	74.6	74.0	72.0
FreeMatch [81]	ViT.	75.0	**75.7**	75.6	75.1	73.9
filtering rate		0.2	0.4	0.6	0.8	baseline only
FreeMatch [81]	Res.	74.5	**74.8**	74.3	73.7	72.0
FreeMatch [81]	ViT.	75.5	75.7	**75.9**	75.5	73.9

Table 8: We ablate a component of our approach with 378 feedback: x_{LD} selection strategy and filtering rate p for bank generation.

find that strategy (ii) achieves the best performance. Therefore, we adopt this strategy for our proposed method.

Further studies, such as extension to a TTA scenario, combining with SFDA methods and different feedback configurations, are presented in Appendix C.

6 Conclusion & Discussion

User feedback can play an integral part in adapting the practical ML product to the target environment. However, we have shown that naive adaptation using existing SemiSDA methods led to undesirable adaptation results. We explained this through the lens of Negatively-Biased Feedback (NBF). In this paper, we uncovered the unexpected results of NBF and presented a scalable solution, *Retrieval Latent Defending*. This method prevents the mini-batch from becoming overly dependent on labeled samples that may have a biased distribution within the overall target distribution. Under the diverse DA benchmarks, from the simulation study to the medical imaging task, we demonstrated the practical problem caused by NBF and the effectiveness of our approach by combining it with multiple SemiSDA baselines. We hope our efforts will inspire future DA works on leveraging user feedback to improve an ML model in the deployment environment.

Broader impact. The proposed setup assumes that an ML product obtains feedback as a form of annotations (*i.e.*, labeled data). In some cases, users can provide feedback in different forms, like thumbs up & down and rating of model prediction, or noise feedback whose information is different from the ground truth. Further research considering these points will pave the way for developing safer and more reliable adapting strategies.

Acknowledgment. We sincerely appreciate the abundant support provided by Lunit Inc., and we would like to thank Donggeun Yoo, Seonwook Park, and Sérgio Pereira for their valuable feedback.

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: International Joint Conference on Neural Networks (IJCNN) (2020) 6, 9, 11, 23, 24, 25, 26, 29, 30
2. Basak, H., Yin, Z.: Semi-supervised domain adaptive medical image segmentation through consistency regularized disentangled contrastive learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2023) 4
3. Baumeister, R.F., Bratslavsky, E., Finkenauer, C., Vohs, K.D.: Bad is stronger than good. Review of general psychology (2001) 2, 5
4. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C.: Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In: ICLR (2020) 4
5. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. In: NeurIPS (2019) 4
6. Berthelot, D., Roelofs, R., Sohn, K., Carlini, N., Kurakin, A.: Adamatch: A unified approach to semi-supervised learning and domain adaptation. In: ICLR (2022) 1, 4, 7, 8, 9, 10, 11, 12, 13, 22, 28, 30, 32, 33, 34
7. Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern recognition (1997) 11
8. Chen, D., Wang, D., Darrell, T., Ebrahimi, S.: Contrastive test-time adaptation. In: CVPR (2022) 3, 9, 10, 24, 25, 27, 28, 30, 35
9. Chen, H., Tao, R., Fan, Y., Wang, Y., Wang, J., Schiele, B., Xie, X., Raj, B., Savvides, M.: Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. In: ICLR (2023) 4
10. Chen, W., Lin, L., Yang, S., Xie, D., Pu, S., Zhuang, Y.: Self-supervised noisy label learning for source-free unsupervised domain adaptation. In: IROS (2022) 29
11. Chen, X., Zhao, Z., Zhang, Y., Duan, M., Qi, D., Zhao, H.: Focalclick: Towards practical interactive image segmentation. In: CVPR (2022) 25
12. Chen, Y., Tan, X., Zhao, B., Chen, Z., Song, R., Liang, J., Lu, X.: Boosting semi-supervised learning by exploiting all unlabeled data. In: CVPR (2023) 4
13. Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M., et al.: Torchxrayvision: A library of chest x-ray datasets and models. In: International Conference on Medical Imaging with Deep Learning (2022) 10, 11, 28
14. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016) 12, 29
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 1, 10, 32, 33, 35
16. Fini, E., Astolfi, P., Alahari, K., Alameda-Pineda, X., Mairal, J., Nabi, M., Ricci, E.: Semi-supervised learning made simple with self-supervised clustering. In: CVPR (2023) 4

17. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML (2015) 1, 3
18. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The journal of machine learning research (2016) 3
19. Gong, T., Jeong, J., Kim, T., Kim, Y., Shin, J., Lee, S.J.: Robust continual test-time adaptation: Instance-aware bn and prediction-balanced memory. In: NeurIPS (2023) 24
20. Harada, S., Bise, R., Araki, K., Yoshizawa, A., Terada, K., Kurata, M., Nakajima, N., Abe, H., Ushiku, T., Uchida, S.: Cluster-guided semi-supervised domain adaptation for imbalanced medical image classification. arXiv preprint arXiv:2303.01283 (2023) 4
21. Hartigan, J.A., Wong, M.A.: Algorithm as 136: A k-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics) (1979) 13, 27
22. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020) 27
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 1, 9, 10, 32, 33, 35
24. Higuchi, Y., Moritz, N., Roux, J.L., Hori, T.: Momentum pseudo-labeling for semi-supervised speech recognition. In: Interspeech (2021) 4
25. Huang, D., Li, J., Chen, W., Huang, J., Chai, Z., Li, G.: Divide and adapt: Active domain adaptation via customized learning. In: CVPR (2023) 4, 21, 22, 29
26. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017) 10, 11, 28
27. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019) 9, 11
28. Khurana, A., Paul, S., Rai, P., Biswas, S., Aggarwal, G.: Sita: Single image test-time adaptation. arXiv preprint arXiv:2112.02355 (2021) 13
29. Kim, J., Hur, Y., Park, S., Yang, E., Hwang, S.J., Shin, J.: Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In: NeurIPS (2020) 22, 23
30. Knox, W.B., Stone, P.: Tamer: Training an agent manually via evaluative reinforcement. In: IEEE international conference on development and learning (2008) 25
31. Lee, H., Shin, S., Kim, H.: Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning. In: NeurIPS (2021) 10, 22
32. Lenga, M., Schulz, H., Saalbach, A.: Continual learning for domain adaptation in chest x-ray classification. In: Medical Imaging with Deep Learning (2020) 10, 11
33. Li, J., Li, G., Shi, Y., Yu, Y.: Cross-domain adaptive clustering for semi-supervised domain adaptation. In: CVPR (2021) 4, 8, 10, 32, 33, 34
34. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: ICML (2020) 1, 3, 4, 10, 26, 27, 35
35. Litrico, M., Del Bue, A., Morerio, P.: Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In: CVPR (2023) 3, 26, 27, 29, 30, 35
36. Liu, X., Xing, F., Shusharina, N., Lim, R., Jay Kuo, C.C., El Fakhri, G., Woo, J.: Act: Semi-supervised domain-adaptive medical image segmentation with asym-

metric co-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2022) 4
37. Liu, X., Yoo, C., Xing, F., Oh, H., El Fakhri, G., Kang, J.W., Woo, J., et al.: Deep unsupervised domain adaptation: A review of recent advances and perspectives. APSIPA Transactions on Signal and Information Processing (2022) 3
38. Liu, Y., Zhang, W., Wang, J.: Source-free domain adaptation for semantic segmentation. In: CVPR (2021) 3
39. Liu, Y., Kothari, P., van Delft, B., Bellot-Gurlet, B., Mordan, T., Alahi, A.: Ttt++: When does self-supervised test-time training fail or thrive? In: NeurIPS (2021) 24, 30
40. Liu, Z., Miao, Z., Pan, X., Zhan, X., Lin, D., Yu, S.X., Gong, B.: Open compound domain adaptation. In: CVPR (2020) 13
41. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: ICML (2017) 3
42. MacGlashan, J., Ho, M.K., Loftin, R., Peng, B., Wang, G., Roberts, D.L., Taylor, M.E., Littman, M.L.: Interactive learning from policy-dependent human feedback. In: International conference on machine learning (2017) 25
43. Madani, A., Moradi, M., Karargyris, A., Syeda-Mahmood, T.: Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In: International symposium on biomedical imaging (2018) 4
44. Mahapatra, D., Korevaar, S., Bozorgtabar, B., Tennakoon, R.: Unsupervised domain adaptation using feature disentanglement and gcns for medical image classification. In: ECCV (2022) 10, 11
45. Mei, K., Zhu, C., Zou, J., Zhang, S.: Instance adaptive self-training for unsupervised domain adaptation. In: ECCV (2020) 3
46. Mei, K., Zhu, C., Zou, J., Zhang, S.: Instance adaptive self-training for unsupervised domain adaptation. In: ECCV (2020) 12, 29
47. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: ICML (2022) 24, 30
48. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. In: ICLR (2023) 13
49. Oh, Y., Kim, D.J., Kweon, I.S.: Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. In: CVPR (2022) 10, 22
50. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. In: NeurIPS (2022) 25
51. Pan, Y., Yao, T., Li, Y., Wang, Y., Ngo, C.W., Mei, T.: Transferrable prototypical networks for unsupervised domain adaptation. In: CVPR (2019) 3
52. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, e.a.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019), https://github.com/pytorch/vision/blob/main/torchvision/models/resnet.py 28
53. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research (2011) 5, 25, 26
54. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: ICCV (2019) 9

55. Prabhu, V., Chandrasekaran, A., Saenko, K., Hoffman, J.: Active domain adaptation via clustering uncertainty-weighted embeddings. In: ICCV (2021) 4, 21, 22, 29
56. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: ECCV (2016) 12, 29
57. Rozin, P., Royzman, E.B.: Negativity bias, negativity dominance, and contagion. Personality and social psychology review (2001) 2, 5
58. Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K.: Semi-supervised domain adaptation via minimax entropy. In: ICCV (2019) 1, 3, 4, 8, 9, 10, 28, 32, 33, 34
59. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR (2018) 3
60. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: NeurIPS (2016) 4
61. Schulman, J., Zoph, B., Kim, C., Hilton, J., Menick, J., Weng, J., Uribe, J.F.C., Fedus, L., Metz, L., Pokorny, M., et al.: Chatgpt: Optimizing language models for dialogue. OpenAI blog (2022) 25
62. Shen, Y., Yun, H., Lipton, Z.C., Kronrod, Y., Anandkumar, A.: Deep active learning for named entity recognition. In: ICLR (2017) 12, 22
63. Shin, I., Kim, D.J., Cho, J.W., Woo, S., Park, K., Kweon, I.S.: Labor: Labeling only if required for domain adaptive semantic segmentation. In: ICCV (2021) 12, 29
64. Sofiiuk, K., Petrov, I., Barinova, O., Konushin, A.: f-brs: Rethinking backpropagating refinement for interactive segmentation. In: CVPR (2020) 25
65. Sofiiuk, K., Petrov, I.A., Konushin, A.: Reviving iterative training with mask guidance for interactive segmentation. In: IEEE International Conference on Image Processing (ICIP) (2022) 25
66. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: NeurIPS (2020) 1, 4, 8, 10, 28, 32, 33
67. Song, J., Lee, J., Kweon, I.S., Choi, S.: Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In: CVPR (2023) 1, 13, 24, 30
68. Song, J., Park, K., Shin, I., Woo, S., Zhang, C., Kweon, I.S.: Test-time adaptation in the dynamic world with compound domain knowledge management. IEEE Robotics and Automation Letters (2023) 24
69. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F.: Learning to summarize with human feedback. In: NeurIPS (2020) 25
70. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: ECCV Workshops (2016) 1, 3
71. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 25
72. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR (2018) 1, 3, 12
73. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR (2018) 3
74. Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine learning (2020) 4

75. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: CVPR (2017) 9
76. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR (2019) 3
77. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: ICLR (2021) 3, 24, 30
78. Wang, M., Deng, W.: Deep visual domain adaptation: A survey. Neurocomputing (2018) 1
79. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: CVPR (2022) 30
80. Wang, Y., Chen, H., Fan, Y., Sun, W., Tao, R., Hou, W., Wang, R., Yang, L., Zhou, Z., Guo, L.Z., et al.: Usb: A unified semi-supervised learning benchmark for classification. In: NeurIPS (2022) 10
81. Wang, Y., Chen, H., Heng, Q., Hou, W., Fan, Y., Wu, Z., Wang, J., Savvides, M., Shinozaki, T., Raj, B., et al.: Freematch: Self-adaptive thresholding for semi-supervised learning. In: ICLR (2023) 4, 7, 8, 10, 12, 13, 14, 24, 25, 27, 28, 30, 32, 33, 34
82. Warnell, G., Waytowich, N., Lawhern, V., Stone, P.: Deep tamer: Interactive agent shaping in high-dimensional state spaces. In: AAAI (2018) 25
83. Wei, C., Sohn, K., Mellina, C., Yuille, A., Yang, F.: Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In: CVPR (2021) 10, 22, 23, 24, 29
84. Wirth, C., Akrour, R., Neumann, G., Fürnkranz, J., et al.: A survey of preference-based reinforcement learning methods. Journal of Machine Learning Research (2017) 25
85. Xie, B., Yuan, L., Li, S., Liu, C.H., Cheng, X.: Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In: CVPR (2022) 12, 29
86. Xie, M., Li, Y., Wang, Y., Luo, Z., Gan, Z., Sun, Z., Chi, M., Wang, C., Wang, P.: Learning distinctive margin toward active domain adaptation. In: CVPR (2022) 4, 21
87. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.: Unsupervised data augmentation for consistency training. In: NeurIPS (2020) 4, 8, 10, 32, 33, 34
88. Xu, Y., Shang, L., Ye, J., Qian, Q., Li, Y.F., Sun, B., Li, H., Jin, R.: Dash: Semi-supervised learning with dynamic thresholding. In: ICML (2021) 4
89. Yan, Z., Wu, Y., Li, G., Qin, Y., Han, X., Cui, S.: Multi-level consistency learning for semi-supervised domain adaptation. In: IJCAI (2022) 4
90. Yang, L., Wang, Y., Gao, M., Shrivastava, A., Weinberger, K.Q., Chao, W.L., Lim, S.N.: Deep co-training with task decomposition for semi-supervised domain adaptation. In: ICCV (2021) 4
91. Yang, S., Jui, S., van de Weijer, J., et al.: Attracting and dispersing: A simple approach for source-free domain adaptation. In: NeurIPS (2022) 3
92. Yang, S., Wang, Y., Van De Weijer, J., Herranz, L., Jui, S.: Generalized source-free domain adaptation. In: ICCV (2021) 1, 3
93. Yang, S., van de Weijer, J., Herranz, L., Jui, S., et al.: Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In: NeurIPS (2021) 3, 27, 28, 35
94. Yu, Y.C., Lin, H.T.: Semi-supervised domain adaptation with source label adaptation. In: CVPR (2023) 4, 9, 10

95. Yu, Z., Li, J., Du, Z., Zhu, L., Shen, H.T.: A comprehensive survey on source-free domain adaptation. arXiv preprint arXiv:2302.11803 (2023) 3
96. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T.: Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In: NeurIPS (2021) 7, 8, 10, 13, 32, 33, 34
97. Zhang, C., Miech, A., Shen, J., Alayrac, J.B., Luc, P.: Making the most of what you have: Adapting pre-trained visual language models in the low-data regime. arXiv preprint arXiv:2305.02297 (2023) 3
98. Zhang, W., Ouyang, W., Li, W., Xu, D.: Collaborative and adversarial network for unsupervised domain adaptation. In: CVPR (2018) 3
99. Zhang, Y., Zhang, H., Deng, B., Li, S., Jia, K., Zhang, L.: Semi-supervised models are strong unsupervised domain adaptation learners. arXiv preprint arXiv:2106.00417 (2021) 1, 4, 8
100. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: ICCV (2017) 13
101. Zhang, Y., Wang, Z., He, W.: Class relationship embedded learning for source-free unsupervised domain adaptation. In: CVPR (2023) 3