SOURCE-FREE UNSUPERVISED DOMAIN ADAPTATION WITH HYPOTHESIS CONSOLIDATION OF PREDICTION RATIONALE

Anonymous authors
Paper under double-blind review

ABSTRACT

Source-Free Unsupervised Domain Adaptation (SFUDA) is a challenging task where a model needs to be adapted to a new domain without access to target domain labels or source domain data. The primary difficulty in this task is that the model's predictions may be inaccurate, and using these inaccurate predictions for model adaptation can lead to misleading results. To address this issue, this paper proposes a novel approach that considers multiple prediction hypotheses for each sample and investigates the rationale behind each hypothesis. By consolidating these hypothesis rationales, we identify the most likely correct hypotheses, which we then use as a pseudo-labeled set to support a semi-supervised learning procedure for model adaptation. To achieve the optimal performance, we propose a three-step adaptation process: model pre-adaptation, hypothesis consolidation, and semi-supervised learning. Extensive experimental results demonstrate that our approach achieves state-of-the-art performance in the SFUDA task and can be easily integrated into existing approaches to improve their performance.

1 INTRODUCTION

The success of deep learning models in visual tasks is largely dependent on whether the training and testing data share similar distributions He et al. (2016); Liang et al. (2020b). However, when the distribution of the testing data differs significantly from that of the training data, also known as domain shift, the performance of these models can decrease substantially Tzeng et al. (2017); Peng et al. (2019). To mitigate the effects of domain shift and reduce the need for data annotations, Unsupervised Domain Adaptation (UDA) techniques have been developed to transfer knowledge from annotated source domains to new but related target domains without requiring annotations in the target domain Hoffman et al. (2018); Long et al. (2018); Dai et al. (2020); Feng et al. (2021); Mei et al. (2020). However, most UDA-based methods rely on access to labeled source domain data during adaptation, such an access may not always be feasible due to privacy concerns. As a result, Source-Free Unsupervised Domain Adaptation (SFUDA) Liang et al. (2020a); Yang et al. (2021b;a); Chen et al. (2022); Yang et al. (2022); Zhang et al. (2022); Karim et al. (2023) gains much attention recently, which only requires a pre-trained model from the source domain and unlabeled data from the target domain.

The main challenge in SFUDA research is how to generate supervision solely from unlabeled data. The current approaches in SFUDA research primarily focus on either generating pseudo-labels Liang et al. (2020a); Yang et al. (2021b;a); Litrico et al. (2023) or conducting unsupervised feature learning Huang et al. (2021); Chen et al. (2022); Zhang et al. (2022); Karim et al. (2023); Litrico et al. (2023) to address this issue. To generate reliable pseudo-labels, existing methods Liang et al. (2020a); Yang et al. (2021b;a) often utilize the distribution of the target domain data to refine the initial predictions from the source domain, i.e., via clustering Liang et al. (2020a) or using the predictions of neighboring samples Yang et al. (2021a); Litrico et al. (2023). On the other hand, unsupervised feature learning, such as contrastive learning, is often employed as an auxiliary task to encourage the features to adapt to the target domain Huang et al. (2021); Chen et al. (2022); Zhang et al. (2022); Karim et al. (2023); Litrico et al. (2023).

In our study, we propose a novel approach to tackle the challenge of SFUDA. Our strategy involves deferring the utilization of label predictions to update the model in the early stages and carefully selecting the most reliable predictions to construct a pseudo-labeled set. The key innovation of our approach lies in considering multiple prediction hypotheses for each sample, accommodating the possibility of multiple potential labels for each data point. We treat each label assignment as a hypothesis and delve into the rationale and supporting evidence behind each prediction. We utilize a representation derived from GradCAM Selvaraju et al. (2017) to encode the rationale for predicting an instance to a hypothetical label. Our methodology is inspired by the belief that assessing the correctness of a prediction can be more reliable by analyzing the reasoning behind a particular prediction, rather than solely relying on prediction probabilities. Subsequently, we develop a consolidation method to determine the most trustworthy hypothesis and utilize it as the labeled dataset in a semi-supervised learning framework. By employing this technique, we effectively transform the SFUDA problem into a conventional semi-supervised learning problem.

Concretely, our approach consists of three key steps: model pre-adaptation, hypothesis consolidation, and semi-supervised learning. We have empirically observed that pre-adapting the model can enhance the effectiveness of the second step. To accomplish this, we introduce a straightforward objective that encourages prediction smoothness from the network. In the final step, we leverage the widely-used FixMatch Sohn et al. (2020) algorithm as our chosen semi-supervised learning method. Through extensive experimentation, we demonstrated the clear advantages of our approach over existing methods in the SFUDA domain and show that the proposed method can be easily integrated into existing approaches to bring improvement.

2 Related Work

UDA. Unsupervised domain adaptation aims to transfer knowledge learned from a labeled source domain to an unlabeled target domain. Various approaches have been proposed to address this task, including discrepancy minimization Tzeng et al. (2014); Ganin & Lempitsky (2015); Long et al. (2015), adversarial learning Hoffman et al. (2018); Long et al. (2018); Tzeng et al. (2017); Vu et al. (2019), and contrastive learning Dai et al. (2020); Kang et al. (2019). Recently, self-training using labeled source data and pseudo-labeled target data has emerged as a prominent approach in unsupervised domain adaptation (UDA) research Feng et al. (2021); Mei et al. (2020); Xie et al. (2020); Yu et al. (2021); Zou et al. (2018). However, these methods typically rely on access to the source data, making them inapplicable when source data is unavailable.

SFUDA. Source-free unsupervised domain adaptation involves adapting a pre-trained model from a source domain to a target domain without access to source data+labels or target labels. Existing SFUDA methods can be broadly categorized into two classes: i) Label Refinement: Methods such as SHOT Liang et al. (2020a), G-SFDA Yang et al. (2021b), NRC Yang et al. (2021a), and GPL Litrico et al. (2023) focus on refining pseudo labels. SHOT generates pseudo labels using centroids obtained in an unsupervised manner. G-SFDA, NRC, and GPL refine pseudo labels through consistent predictions and nearest neighbor knowledge aggregation from local neighboring samples. ii) Contrastive Feature Learning: Approaches like HCL Huang et al. (2021), C-SFDA Karim et al. (2023), AdaContrast Chen et al. (2022), GPL Litrico et al. (2023), and DaC Zhang et al. (2022). HCL and C-SFDA use a contrastive loss similar to moco He et al. (2016), where positive pairs consist of augmented query samples and negatives are other samples. AdaContrast and GPL exclude same-class negative pairs based on pseudo labels. DaC divides the target data into source-like and target-specific samples, computes source-like class centroids, and generates negative pairs using these centroids. These methods aim to tackle SFUDA by refining pseudo labels or leveraging contrastive feature learning, demonstrating the potential of different strategies in addressing the challenges of adapting models without access to labeled source data or target label.

3 Method

In the source-free unsupervised domain adaptation (SFUDA) setting, only pretrained source models and unlabeled data in the target domain are given. The task is to adapt the model to the target domain by using unlabeled target data only. Our approach sequentially applies three steps as described in Sec. 3.1), Sec. 3.2) and Sec. 3.3).

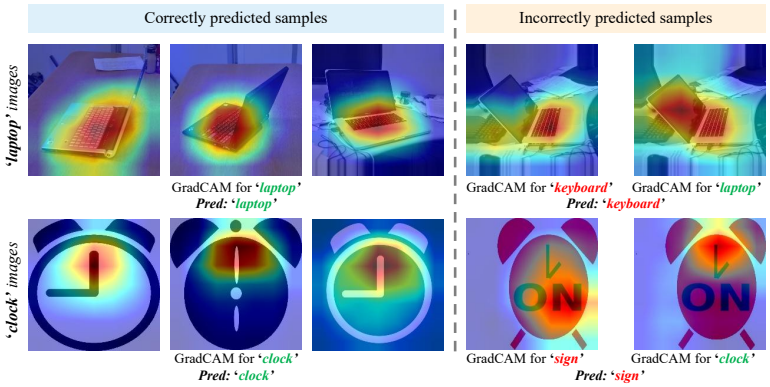


Figure 1: The visualizations illustrate the GradCAM Selvaraju et al. (2017) for predicting the image to a specific class. In the right-half section, it can be observed that even though the prediction is incorrect, the obtained rationale (region highlighted in the GradCAM) based on the correct label remains reasonable and resembles the rationale of the corresponding class depicted in the left-half section.

3.1 Model Pre-adaptation via Encouraging Smooth Prediction

The first step of our approach is to make an initial adaptation to reduce the domain gap. We empirically find such a step can be beneficial for the following steps. We develop a pre-adaptation strategy by encouraging a smooth prediction on the data manifold. [1] Specifically, we create a memory $Q \in \mathbb{R}^{N_q \times d}$ to store N_q randomly sampled image features and update it after each batch training (we choose N_q equals the number of target samples in the dataset). Then for each target sample x_i, we find the z-nearest neighbor $\mathcal{NN}(x_i)$ and z-samples $\mathcal{FN}(x_i)$ that are furthest to x_i based on the Euclidean distance between the image feature of x_i and features in Q (we choose $z = 3$ in our implementation). Then we optimize the following objective:

$$\mathcal{L}_{PA} = \mathcal{L}_{SM} + \lambda \mathcal{L}_{FAR} = \sum_{i=1}^{N_B} \sum_{x'_j \in \mathcal{NN}(x_i)} KL(p(x_i), p(x'_j)) + \lambda \sum_{i=1}^{N_B} \sum_{x'_j \in \mathcal{FN}(x_i)} p(x_i)^\top p(x'_j), \tag{1}$$

where KL represents Kullback-Leibler divergence and p denotes the posterior probability predicted from the source model. N_B is the number of samples within a mini-batch. The first term is used to ensure similar samples have similar predictions. However, using the first term alone may lead to a trivial solution that assigns identical prediction for every instance. Thus we use the second term to counter-act it as it ensures that the least similar samples should have divergent posterior probabilities, i.e., the inner product between posterior should close to zero.

3.2 Hypothesis Consolidation from Prediction Rationale

After pre-adaptation, the model generally exhibits improved adaptation to the target domain. However, there may still be instances where the model produces incorrect predictions, making it challenging to rectify misclassifications solely based on predicted posterior probabilities. Therefore, in the second step, we explore a more robust methodology for analyzing predictions.

We begin by considering multiple prediction hypotheses for each individual instance. Specifically, for each instance, we consider the top $\tilde{k}$ classes with the highest posterior probabilities as potential prediction hypotheses, denoted as (x_i, y^h_{ik}), $k \in$ top $\tilde{k}$. In other words, we acknowledge the correct class label could exist within one of these top $\tilde{k}$ classes, even though we do not know which one.

To further analyze each hypothesis (x_i, y^h_{ik}), we calculate the GradCAM Selvaraju et al. (2017) to identify the regions that contribute to supporting the prediction for y^h_{ik}, resulting in a representation

[1] Other pre-adaptation approaches may also work, such as the method in Liang et al. (2020a) Yang et al. (2022), please refer to Sec. 4.4 for more experimental evidence.

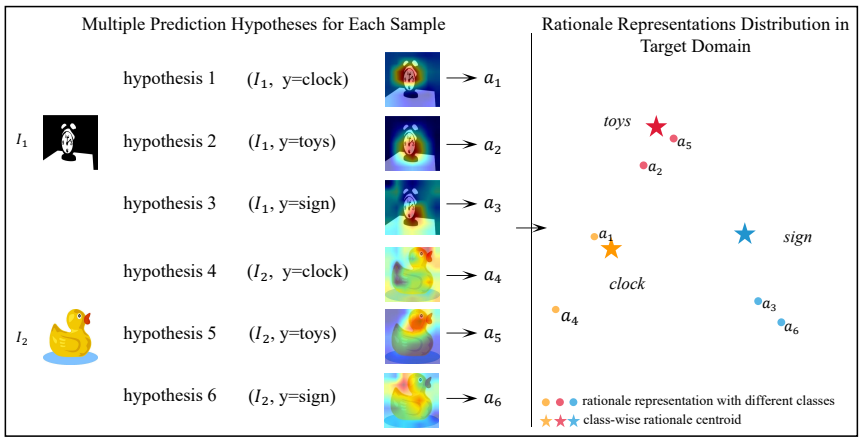


Figure 2: In our method, we generate multiple prediction hypotheses based on the posterior probability of the current model. An image I and its hypothetical label form a hypothesis, for example, (I, $y =$ clock). For each hypothesis, GradCAM is calculated based on the hypothetical label, resulting in the corresponding rationale representation a. Subsequently, we calculate the centroid for the rationale representation of each class.

called the rationale representation a_{ik}. This rationale representation encodes the evidence supporting the corresponding hypothesis. Drawing inspiration from prior work Shu et al. (2022; 2023), we formally calculate a_{ik} using the following equation:

$$a_{ik} = \frac{1}{HW} \sum_{m=1}^{H} \sum_{n=1}^{W} \left(\left[\frac{\partial logit(y_{ik}^h)}{\partial [\phi(x_i)]_{m,n}}^\top [\phi(x_i)]_{m,n} \right]_+ \cdot [\phi(x_i)]_{m,n} \right) \in \mathbb{R}^{d'}, \tag{2}$$

where $\phi(x_i) \in \mathbb{R}^{H \times W \times d'}$ is the feature map of the last convolutional layer of the network with H height, W width, and d' channels. $[\phi(x_i)]_{m,n} \in \mathbb{R}^{d'}$ is the feature vector located at the (m, n)-th grid. $logit(y_{ik}^h)$ is the logit for class y_{ik}^h, $[\cdot]_+ = max(\cdot, 0)$. $\left[\frac{\partial logit(y_{ik}^h)}{\partial [\phi(x_i)]_{m,n}}^\top [\phi(x_i)]_{m,n} \right]_+$ is equivalent to GradCAM value at the (m, n)-th grid. Essentially, the calculation of a_{ik} performs weighted average pooling over $\phi(x_i)$ according to the GradCAM. Figure 1 shows the GradCAM calculated from different hypotheses for the same image. Upon observation, we notice that even if the ground-truth class is not ranked as the top prediction by the model, its associated rationale remains reasonable and similar to the common rationale patterns for the corresponding class. This inspires us to leverage this observation to analyze the model's current predictions. For example, if an instance has a prediction hypothesis that exhibits a rationale similar to the corresponding class's common rationale but is not ranked as the top prediction, then the top prediction may not be correct.

Formally, we calculate the class-wise rationale centroid as the average rationale representation from each hypothetical class, representing the common rationale for each class:

$$\bar{a}_c = \frac{\sum_{ik} \mathbb{1}(y_{ik}^h = c) a_{ik}}{\sum_{ik} \mathbb{1}(y_{ik}^h = c)}, \tag{3}$$

where c represents a class and $\mathbb{1}(y_{ik}^h = c) = 1$ if $k = c$. The idea of using multiple hypotheses with the rationale representation is illustrated in Figure 2.

Next, we generate a ranking index r_{ik} for each prediction hypothesis (x_i, a_{ik}, y_{ik}^h) by ranking the Euclidean distance between a_{ik} and its corresponding rationale centroid $\bar{a}_{y_{ik}^h}$, i.e., the centroid for class y_{ik}^h, in the ascending order. For each instance x_i, we obtain $\tilde{k}$ ranking indices r_{ik}, $k \in$ top $\tilde{k}$ classes, one for each hypothesis. Then, a hypothesis $\{x_i, y_{ik'}\}$ is considered reliable if it satisfies the following two conditions: (1) $r_{ik'} < \tau_1$, indicating the rationale for $\{x_i, y_{ik'}\}$ is typical as its rationale representation is close to the rationale centroid. (2) $r_{ij} > \tau_2 \ \ \forall j \neq k'$, where $\tau_2 > \tau_1$

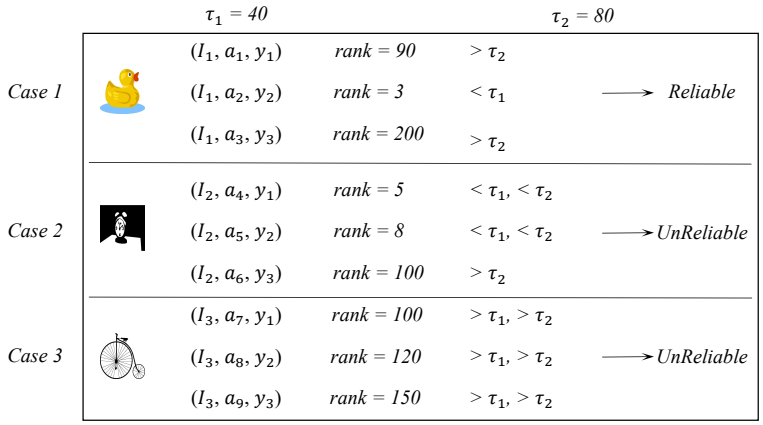


Figure 3: These examples demonstrate the generation of reliable hypotheses. In Case 1, the rank ID of the second hypothesis derived from the image is lower than τ_1, while all other hypotheses from the same image have ranks larger than τ_2. Consequently, the second hypothesis of I_1 is selected as a reliable hypothesis. In Case 2, no hypothesis is selected because it has two hypotheses with rank IDs less than τ_2, indicating a conflict between those hypotheses. Similarly, Case 3 is not selected because none of its hypotheses has rank IDs lower than τ_1.

are two predefined ranking thresholds. The second condition ensures that there is no conflicting hypotheses, i.e., no other hypothesis is likely to be true for the same instance as their rationale appears to be unusual.

With those criteria, we can collect a set of reliable hypotheses $\mathcal{P}$ as samples with their corresponding hypothetical labels. Representative examples of this procedure are depicted in Figure 3. It is important to note that in the second step, we aim to select the most reliable hypothesis rather than correcting hypotheses. This is because we believe that the task of correcting predictions or hypotheses can be better accomplished through the use of semi-supervised learning, which allows for the gradual propagation of pseudo-labels.

By focusing on identifying the most reliable hypothesis based on the proximity of the rationale representation to the rationale centroid and the absence of conflicting rationales, we can create a high-quality set of pseudo-labeled samples (see Appendix C and D). These pseudo-labels can then be used in a semi-supervised learning framework to refine the model's predictions and gradually improve its performance.

3.3 Semi-Supervised Learning

After completing the second step of hypothesis consolidation, we obtain a reliable pseudo-label set $\mathcal{P}$, while the remaining samples are treated as the unlabeled set $\mathcal{U}$. At this stage, we are ready to apply a semi-supervised algorithm to perform the final step of adaptation. For this purpose, we utilize one of the state-of-the-art semi-supervised methods, FixMatch Sohn et al. (2020), which combines consistency regularization and pseudo-labeling to address this task.

Specifically, we start by sampling a labeled mini-batch $\mathcal{B}_l$ from the reliable pseudo-label set $\mathcal{P}$ and an unlabeled batch $\mathcal{B}_u$ from the unlabeled set $\mathcal{U}$. We then optimize the following objective function using these batches:

$$\mathcal{L}_{FM} = \sum_{x_b \in \mathcal{B}_l} CE(\hat{y}_b, p(\mathcal{A}_w(x_b))) + \sum_{x_u \in \mathcal{B}_u} \mathbb{1}(\max(p(\mathcal{A}_w(x_u))) \geq \tau) CE(\hat{y}_u, p(\mathcal{A}_s(x_u))), \quad (4)$$

where $\hat{y}_u = \arg\max_c p(y = c|\mathcal{A}_w(x_u))$. $\mathcal{A}_w(\cdot)$ and $\mathcal{A}_s(\cdot)$ are the weakly-augmented and strongly-augmented operations, respectively. τ is the threshold defined in FixMatch to identify reliable pseudo-label (we set the same with FixMatch as 0.95), and CE is the cross-entropy between two probability distributions.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. Office-Home Venkateswara et al. (2017) consists of 15,500 images categorized into 65 classes. It includes four distinct domains: Real-world (Rw), Clipart (Cl), Art (Ar), and Product (Pr). To evaluate the proposed method, researchers perform 12 transfer tasks on this dataset, involving adapting models across the four domains. The evaluation reports each domain shift Top-1 and the average Top-1 accuracy. Originally, the **DomainNet** dataset Peng et al. (2019) consisted of over 500,000 images, including six domains and 345 classes. For our evaluation, we follow the approach described in Saito et al. (2019) and focus on four domains: Real World (Rw), Sketch (Sk), Clipart (Cl), and Painting (Pt), resulting in **DomainNet-126**. We assess our proposed method on seven domain shifts within these four domains. **VisDA-C** Peng et al. (2017) contains 152,000 synthetic images from the source domain and 55,000 real object images from the target domain. It consists of 12 object classes, and there is a significant synthetic-to-real domain gap between the two domains. Our evaluation reports per-class Top-1 accuracies, as well as the average Top-1 accuracy on this dataset. The **implementation details** of our method can be found in the Appendix A.

4.2 COMPARISON WITH STATE-OF-THE-ARTS

Table 1: Accuracy (%) on medium-sized **Office-Home** dataset (ResNet-50). "SF" denotes source-free.

Method	SF	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
ResNet-50 He et al. (2016)	×	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
GSDA Hu et al. (2020)	×	61.3	76.1	79.4	65.4	73.3	74.3	65.0	53.0	80.0	72.2	60.6	83.1	70.3
RSDA Gu et al. (2020)	×	53.3	77.7	81.3	66.4	74.0	76.5	67.9	53.0	82.0	75.8	57.8	85.4	70.9
SRDC Tang et al. (2020)	×	52.3	76.3	81.0	69.5	76.2	78.0	**68.7**	53.8	81.7	76.3	57.1	85.0	71.3
FixBi Na et al. (2021)	×	58.1	77.3	80.4	67.7	79.5	78.1	65.8	57.9	81.7	76.4	62.9	**86.7**	72.7
G-SFDA Yang et al. (2021b)	✓	57.9	78.6	81.0	66.7	77.2	77.2	65.6	56.0	82.2	72.0	57.8	83.4	71.3
SHOT Liang et al. (2020a)	✓	56.9	78.1	81.0	67.9	78.4	78.1	67.0	54.6	81.8	73.4	58.1	84.5	71.6
SHOT++ Liang et al. (2021)	✓	57.9	79.7	82.5	68.5	79.6	79.3	68.5	57.0	83.0	73.7	60.7	84.9	73.0
NRC Yang et al. (2021a)	✓	57.7	**80.3**	82.0	68.1	79.8	78.6	65.3	56.4	83.0	71.0	58.6	85.6	72.2
CoWA Lee et al. (2022)	✓	56.9	78.4	81.0	69.1	80.0	79.9	67.7	57.2	82.4	72.8	60.5	84.5	72.5
HCL Huang et al. (2021)	✓	**64.0**	78.6	82.4	64.5	73.1	80.1	64.8	**59.8**	75.3	**78.1**	**69.3**	81.5	72.6
DaC Zhang et al. (2022)	✓	59.1	79.5	81.2	69.3	78.9	79.2	67.4	56.4	82.4	74.0	61.4	84.4	72.8
VMP Jing et al. (2022)	✓	57.9	77.6	82.5	68.6	79.4	**80.6**	68.4	55.6	83.1	75.2	59.6	84.7	72.8
SFDA-DE Ding et al. (2022)	✓	59.7	79.5	82.4	69.7	78.6	79.2	66.1	57.2	82.6	73.9	60.8	85.2	72.9
C-SFDA Karim et al. (2023)	✓	60.3	80.2	**82.9**	69.3	80.1	78.8	67.3	58.1	83.4	73.6	61.3	86.3	73.5
Ours	✓	59.9	79.6	82.7	**70.3**	**81.8**	80.4	68.5	57.8	**83.5**	72.5	59.8	86.0	**73.6**

Table 2: Accuracy (%) on large-scale **DomainNet-126** dataset (ResNet-50). "SF" denotes source-free.

Method	SF	Rw→Cl	Rw→Pt	Pt→Cl	Cl→Sk	Sk→Pt	Rw→Sk	Pt→Rw	Avg.
ResNet-50 He et al. (2016)	×	58.8	62.2	57.7	50.3	52.6	47.3	73.2	57.4
MCC Jin et al. (2020)	×	44.8	65.7	41.9	34.9	47.3	35.3	72.4	48.9
CDAN Long et al. (2018)	×	65.0	64.9	63.7	53.1	63.4	54.5	73.2	62.5
GVB Cui et al. (2020)	×	68.2	69.0	63.2	56.6	63.1	62.2	78.3	65.2
MME Saito et al. (2019)	×	70.0	67.7	69.0	56.3	64.8	61.0	76.0	66.4
TENT Wang et al. (2020)	✓	58.5	65.7	57.9	48.5	52.4	54.0	67.0	57.7
G-SFDA Yang et al. (2021b)	✓	63.4	67.5	62.5	55.3	60.8	58.3	75.2	63.3
NRC Yang et al. (2021a)	✓	67.5	68.0	67.8	57.6	59.3	58.7	74.3	64.7
SHOT Liang et al. (2020a)	✓	67.7	68.4	66.9	60.1	66.1	59.9	80.8	67.1
AdaConstrast Chen et al. (2022)	✓	70.2	69.8	68.6	58.0	65.9	61.5	80.5	67.8
DaC Zhang et al. (2022)*	✓	70.0	68.8	70.9	62.4	66.8	60.3	78.6	68.3
C-SFDAKarim et al. (2023)	✓	70.8	71.1	68.5	62.1	67.4	62.7	80.4	69.0
GPL Litrico et al. (2023)	✓	74.2	70.4	68.8	64.0	67.5	**65.7**	76.5	69.6
Ours	✓	**76.9**	**71.8**	**75.4**	**65.5**	**69.9**	64.6	**83.2**	**72.5**

* This work uses ResNet-34 as backbone.

We compare our proposed method against popular source-present and source-free methods on three benchmark datasets: Office-Home, DomainNet-126, and VisDA-C. We report the Top-1 accuracy, and the results are presented in Table 1 to Table 3. In the Office-Home dataset, as shown in Table 1, our proposed method achieves the best performance in terms of Top-1 average accuracy, which is comparable to the most recent source-free method C-SFDA. Additionally, our method in 3 sub-transfer tasks achieves the highest accuracy (see bold in Table 1) vs. only one sub-transfer task in C-SFDA. For the DomainNet-126 dataset, as demonstrated in Table 2, our proposed method exhibits

Table 3: Accuracy (%) on large-scale **VisDA-C** dataset (ResNet-101). "SF" denotes source-free.

Method	SF	plane	bcyle	bus	car	horse	knife	mcyle	person	plant	sktbrd	train	truck	Avg.
ResNet-101 He et al. (2016)	×	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
MCC Jin et al. (2020)	×	88.7	80.3	80.5	71.5	90.1	93.2	85.0	71.6	89.4	73.8	85.0	36.9	78.8
STAR Lu et al. (2020)	×	95.0	84.0	84.6	73.0	91.6	91.8	85.9	78.4	94.4	84.7	87.0	42.2	82.7
RWOT Xu et al. (2020)	×	95.1	87.4	85.2	58.6	96.2	95.7	90.6	80.0	94.8	90.8	88.4	47.9	84.3
CAN Kang et al. (2019)	×	97.0	87.2	82.5	74.3	**97.8**	96.2	90.8	80.7	96.6	**96.3**	87.5	59.9	87.2
SHOT Liang et al. (2020a)	✓	94.3	88.5	80.1	57.3	93.1	94.9	80.7	80.3	90.5	89.1	86.3	58.2	82.9
DIPE Wang et al. (2022)	✓	95.2	87.6	78.8	55.9	93.9	95.0	84.1	81.7	92.1	88.9	85.4	58.0	83.1
HCL Huang et al. (2021)	✓	93.3	85.4	80.7	68.5	91.0	88.1	86.0	78.6	86.6	88.8	80.0	**74.7**	83.5
A^2Net Xia et al. (2021)	✓	94.0	87.8	85.6	66.8	93.7	95.1	85.8	81.2	91.6	88.2	86.5	56.0	84.3
G-SFDA Yang et al. (2021b)	✓	96.1	88.3	85.5	74.1	97.1	95.4	89.5	79.4	95.4	92.9	89.1	42.6	85.4
NRC Yang et al. (2021a)	✓	96.8	**91.3**	82.4	62.4	96.2	95.9	86.1	80.6	94.8	94.1	90.4	59.7	85.9
SFDA-DE Ding et al. (2022)	✓	95.3	91.2	77.5	72.1	95.7	**97.8**	85.5	86.1	95.5	93.0	86.3	61.6	86.5
AdaContrast Chen et al. (2022)	✓	97.0	84.7	84.0	77.3	96.7	93.8	91.9	84.8	94.3	93.1	**94.1**	49.7	86.8
CoWA Lee et al. (2022)	✓	96.2	89.7	83.9	73.8	96.4	97.4	89.3	**86.8**	94.6	92.1	88.7	53.8	86.9
DaC Zhang et al. (2022)	✓	96.6	86.8	**86.4**	78.4	96.4	96.2	**93.6**	83.8	**96.8**	95.1	89.6	50.0	87.3
BDT Kundu et al. (2022)	✓	-	-	-	-	-	-	-	-	-	-	-	-	87.8
C-SFDA Karim et al. (2023)	✓	97.6	88.8	86.1	72.2	97.2	94.4	92.1	84.7	93.0	90.7	93.1	63.5	87.8
Ours	✓	**98.0**	88.0	**86.4**	**82.3**	**97.8**	96.2	92.1	85.0	95.5	91.7	93.8	56.2	**88.6**

Table 4: Ablation study of the proposed components calculated by average accuracy (%) on the **Office-Home** (O-H), **DomainNet-126** (DN-126) and **VisDA-C** datasets. PA stands for model pre-adaptation (Sec. 3.1), HCPR (Sec. 3.2) stands for hypothesis consolidation from prediction rationale, FM stands for FixMatch techniques (Sec. 3.3).

#	PA	HCPR	FM	O-H	DN-126	VisDA-C
0	×	×	×	60.2	55.6	46.6
1	×	×	✓	64.2	60.6	62.3
2	×	✓	✓	68.6	70.6	85.2
3	✓	×	×	72.1	67.4	86.2
4	✓	✓	×	72.7	69.6	87.5
5	✓	×	✓	72.2	67.5	86.2
6	✓	✓	✓	**73.6**	**72.5**	**88.6**

Table 5: **DomainNet-126** (Pt→Cl) Top-1 accuracy (%) of the proposed method with different number of the prediction hypotheses $\tilde{k}$. We find $\tilde{k} = 4$ yields the optimal results.

$\tilde{k}$	Accuracy
2	73.7
3	74.2
4	**75.4**
5	75.3
6	74.8
10	71.8
20	66.9

significant improvements over all baselines. With an average Top-1 accuracy of 72.5%, our method outperforms the best source-free baseline by nearly 3% and surpasses the best source-present baseline by 6.1%. Moreover, our method achieves the best performance in almost all domain shifts. On the VisDA-C dataset, presented in Table 3, our proposed method outperforms the state-of-the-art method C-SFDA Karim et al. (2023) by 0.8%. Furthermore, our method achieves the best performance in specific classes such as "plane", "bus", "car", and "horse". These results clearly demonstrate the superiority of our proposed method across the evaluated datasets, showcasing its effectiveness in source-free domain adaptation scenarios.

4.3 Ablation Studies

Component-wise analysis.

In this section, we conduct ablation studies to analyze the contribution of each component in our method on three benchmark datasets: Office-Home, DomainNet-126, and VisDA-C. The results are summarized in Table 4. Each component of our methods helps to enhance the performance, in which the HCPR (Hypothesis Consolidation from Prediction Rationale) component makes the most contributions to the promotion of accuracy. Specifically, compared to only using FixMatch, combining both FixMatch and HCPR significantly improves accuracy by 4.4%, 10.0%, and 22.9% on the respective datasets. Additionally, in the case of combining both PA (Pre-Adaptation) and HCPR, we execute PA again following HCPR to integrate the consolidation outcomes from HCPR. This showcases a substantial enhancement in accuracy, with improvements of 0.6%, 2.2%, and 1.3% on the respective datasets compared to solely employing PA. Last but not least, Removing HCPR from the

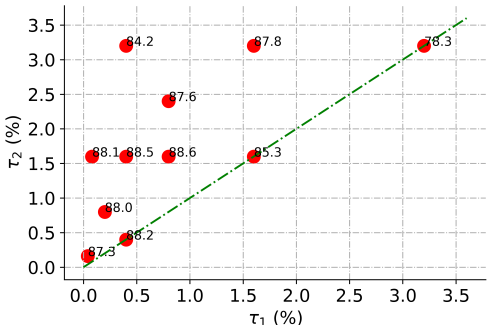


Figure 4: **VisDA-C** average accuracy (%) of the proposed method using different τ_1 and τ_2.

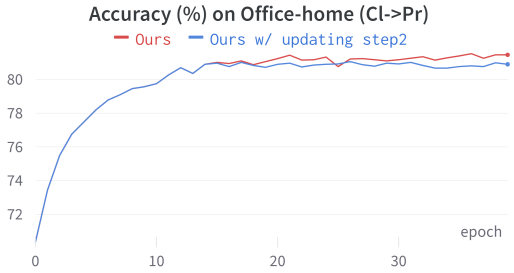


Figure 5: **Office-Home** accuracy of Ours and Ours w/ updating step 2 across varying epochs.

method leads to a performance drop of 1.4%, 5%, and 2% points on Office-Home, DomainNet-126, and VisDA-C, respectively.

Impact of $\tilde{k}$ —the number of prediction hypotheses per instance.

In our method, we choose labels from the top $\tilde{k}$ highest posterior probabilities as the prediction hypothesis. In this section, we investigate the impact of the value of $\tilde{k}$. Table 5 shows the accuracy achieved with different $\tilde{k}$. From the result, we can see that using 2 hypotheses has already led to good performance. When $\tilde{k}$ is very large, the performance drop significantly. As a result, it is recommended to set $\tilde{k}$ in a smaller range, and specifically, choosing 3-6 hypotheses leads to optimal performance.

Impact of the two ranking thresholds τ_1 and τ_2.

To assess the influence of ranking thresholds in our method, we examined the percentage values τ_1 and τ_2 relative to the total number of samples. Specifically, we analyzed their impact on the Top-1 average accuracy on the VisDA-C dataset, as illustrated in Figure 4. Our analysis, depicted in Figure 4, revealed that the proposed method exhibits robustness to the specific values of τ_1 and τ_2.

The benefit of using rationale representations.

To further understand the benefit of using the rationale representation from multiple hypotheses, we explore an alternative method that replaces the proposed second step by using feature centroids rather than rationale centroids. Since the feature is invariant to the prediction hypothesis, only the top predicted class will be considered. More specially, we first generate pseudo-label for each instance and calculate the feature centroid similar to our approach. Then we rank instances based on the Euclidean distances between their features and the corresponding class centroid. The top τ_1 features closest to the class centroid are assigned reliable pseudo labels, while the remaining samples are left for step 3. We refer to this method as "near-centroid selection". Table 6 presents the comparison results on the Office-Home and DomainNet-126 datasets. As seen, while such an approach still leads to improvement over using step 1 and step 3 alone (by cross-referencing Table 4), it is still inferior to the use of HCPR. This clearly demonstrates the benefits of the latter.

Table 6: Average accuracy (%) of our HCPR vs. near centroid collection on the **Office-Home** (O-H) and **DomainNet-126** (DN-126) datasets.

Method	O-H	DN-126
near-centroid selection	72.6	69.6
Ours	**73.6**	**72.5**

Investigation of recursively applying HCPR.

One may wonder if recursively applying HCPR will lead to additional improvement. To this end, we create a variant of our method by alternatively applying step 2 and step 3, hoping that they may mutually enhance each other. We conducted experiments on the Office-Home (Cl→Pr) dataset. The results are depicted in Figure 5, where the red curve represents our method using the second step only once, i.e., the hypothesis consolidation occurs between model pre-adaptation (0-9 epochs) and

Table 7: Accuracy (%) of our method combined with existing SHOT and AaD methods on the **Office-Home**, **VisDA-C** and **DomainNet-126** datasets.

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
SHOT Liang et al. (2020a)	56.9	78.1	81.0	67.9	78.4	78.1	67.0	54.6	81.8	73.4	58.1	84.5	71.6
SHOT+Ours	58.7	**79.5**	82.1	69.6	80.7	**80.0**	**69.1**	56.9	82.3	**74.5**	59.2	85.3	73.2
AaD Yang et al. (2022)	59.3	79.3	82.1	68.9	79.8	79.5	67.2	57.4	83.1	72.1	58.5	85.4	72.7
AaD+Ours	**59.8**	79.4	**82.7**	**70.0**	**81.6**	**80.0**	68.5	**57.6**	**83.2**	72.7	**59.4**	**86.1**	**73.4**

Method	plane	bcyle	bus	car	horse	knife	mcyle	person	plant	sktbrd	train	truck	Avg.
SHOT Liang et al. (2020a)	94.3	88.5	80.1	57.3	93.1	94.9	80.7	80.3	90.5	89.1	86.3	58.2	82.9
SHOT+Ours	97.5	84.6	83.0	74.2	96.5	93.7	92.8	**86.7**	93.5	92.6	89.7	56.9	86.8
AaD Yang et al. (2022)	97.4	**90.5**	80.8	76.2	97.3	**96.1**	89.8	82.9	**95.5**	**93.0**	92.0	**64.0**	88.0
AaD+Ours	**97.8**	87.6	**86.7**	**83.4**	**97.7**	95.4	**94.2**	83.8	94.6	91.2	**92.8**	55.6	**88.4**

Method	Rw→Cl	Rw→Pt	Pt→Cl	Cl→Sk	Sk→Pt	Rw→Sk	Pt→Rw	Avg.
SHOT Liang et al. (2020a)	67.7	68.4	66.9	60.1	66.1	59.9	80.8	67.1
SHOT+Ours	70.5	70.6	72.5	63.6	68.0	61.1	**82.8**	69.9
AaD Yang et al. (2022)	70.6	69.8	69.3	58.5	66.2	60.2	80.2	67.8
AaD+Ours	**75.4**	**71.3**	**75.2**	**64.2**	**68.4**	**63.3**	**82.8**	**71.5**

semi-supervised learning (10-40 epochs). The blue curve represents our method with the second step updated at the 15th, 20th, and 25th epochs. From the results, we observed that recursively applying HCPR does not lead to an improvement as one may expect. We also conduct experiments with HCPR applied recursively to only PA or FixMatch, which can be found in the Appendix F.

4.4 Incorporating the Proposed Method into Existing Approaches

The proposed method can be seamlessly integrated into existing network architectures, such as SHOT Liang et al. (2020a) and AaD Yang et al. (2022). Specifically, we replace the pre-adaptation phase in our first step with SHOT and AaD, resulting in the combined approach referred to as "SHOT+Ours" and "AaD+Ours". The experimental results, as shown in Table 7, demonstrate the superiority of the proposed method integrated into the SHOT and AaD objectives. Across the Office-Home (Avg. $\uparrow 1.6\%$ and $\uparrow 0.7\%$), VisDA-C (Avg. $\uparrow 3.9\%$ and $\uparrow 0.4\%$), and DomainNet-126 (Avg. $\uparrow 2.8\%$ and $\uparrow 3.7\%$) datasets, the integrated approach consistently outperforms the baseline of SHOT and AaD. This indicates that our method complements existing SFUDA baselines and consistently improves their performance by incorporating our approach as a replacement for the model pre-adaptation phase.

5 Limitation and Future Work

The current approach relies on having access to the entire target training set to perform crucial steps like pre-adaptation and identifying the reliable pseudo-labeled set. However, in real-world applications, online adaptation is often more desirable as it doesn't require holding a large number of target examples. As part of our future work, we aim to extend the key idea of this research to the online streaming setting. By doing so, we can develop a methodology that adapts in real time to incoming data, allowing for more efficient and effective adaptation in dynamic environments. This extension will enhance the applicability and practicality of the proposed approach in various domains.

6 Conclusion

In conclusion, this paper introduces a novel approach for Source-Free Unsupervised Domain Adaptation (SFUDA), where a model needs to adapt to a new domain without access to target domain labels or source domain data. By considering multiple prediction hypotheses and analyzing their rationales, the proposed method identifies the most likely correct hypotheses, which are then used as pseudo-labeled data for a semi-supervised learning procedure. The three-step adaptation process, including model pre-adaptation, hypothesis consolidation, and semi-supervised learning, ensures optimal performance. Experimental results demonstrate that the proposed approach achieves state-of-the-art performance in the SFUDA task and can be seamlessly integrated into existing methods to enhance their performance.

REFERENCES

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 295–305, 2022.

Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12455–12464, 2020.

Shuyang Dai, Yu Cheng, Yizhe Zhang, Zhe Gan, Jingjing Liu, and Lawrence Carin. Contrastively smoothed class alignment for unsupervised domain adaptation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7212–7222, 2022.

Hao Feng, Minghao Chen, Jinming Hu, Dong Shen, Haifeng Liu, and Deng Cai. Complementary pseudo labels for unsupervised domain adaptation on person re-identification. *IEEE Transactions on Image Processing*, 30:2898–2907, 2021.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Xiang Gu, Jian Sun, and Zongben Xu. Spherical space domain adaptation with robust pseudo-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9101–9110, 2020.

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*, pp. 770–778, 2016.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pp. 1989–1998. Pmlr, 2018.

Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Unsupervised domain adaptation with hierarchical gradient synchronization. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 4043–4052, 2020.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34:3635–3649, 2021.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pp. 464–480. Springer, 2020.

Mengmeng Jing, Xiantong Zhen, Jingjing Li, and Cees Snoek. Variational model perturbation for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 35: 17173–17187, 2022.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4893–4902, 2019.

Nazmul Karim, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-pang Chiu, Supun Samarasekera, and Nazanin Rahnavard. C-sfda: A curriculum learning aided self-training framework for efficient source free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24120–24131, 2023.

Jogendra Nath Kundu, Akshay R Kulkarni, Suvaansh Bhambri, Deepesh Mehta, Shreyas Anand Kulkarni, Varun Jampani, and Venkatesh Babu Radhakrishnan. Balancing discriminability and transferability for source-free domain adaptation. In *International Conference on Machine Learning*, pp. 11710–11728. PMLR, 2022.

Jonghyun Lee, Dahuin Jung, Junho Yim, and Sungroh Yoon. Confidence score for source-free unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 12365–12377. PMLR, 2022.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 6028–6039. PMLR, 2020a.

Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8602–8617, 2021.

Justin Liang, Namdar Homayounfar, Wei-Chiu Ma, Yuwen Xiong, Rui Hu, and Raquel Urtasun. Polytransform: Deep polygon transformer for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9131–9140, 2020b.

Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7640–7650, 2023.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9111–9120, 2020.

Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pp. 415–430. Springer, 2020.

Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1094–1103, 2021.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8050–8058, 2019.

Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Yangyang Shu, Baosheng Yu, Haiming Xu, and Lingqiao Liu. Improving fine-grained visual recognition in low data regimes via self-boosting attention mechanism. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, pp. 449–465. Springer, 2022.

Yangyang Shu, Anton van den Hengel, and Lingqiao Liu. Learning common rationale to improve self-supervised representation for fine-grained visual recognition problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11392–11401, 2023.

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.

Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8725–8735, 2020.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2517–2526, 2019.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Fan Wang, Zhongyi Han, Yongshun Gong, and Yilong Yin. Exploring domain-invariant parameters for source free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7151–7160, 2022.

Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9010–9019, 2021.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698, 2020.

Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, and Jindong Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4394–4403, 2020.

Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in neural information processing systems*, 34:29393–29405, 2021a.

Shiqi Yang, Yaxing Wang, Joost Van De Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8978–8987, 2021b.

Shiqi Yang, Yaxing Wang, Kai Wang, Shangling Jui, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. In *Advances in Neural Information Processing Systems*, 2022.

Fei Yu, Mo Zhang, Hexin Dong, Sheng Hu, Bin Dong, and Li Zhang. Dast: Unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10754–10762, 2021.

Ziyi Zhang, Weikai Chen, Hui Cheng, Zhen Li, Siyuan Li, Liang Lin, and Guanbin Li. Divide and contrast: Source-free domain adaptation via adaptive contrastive learning. *arXiv preprint arXiv:2211.06612*, 2022.

Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 289–305, 2018.

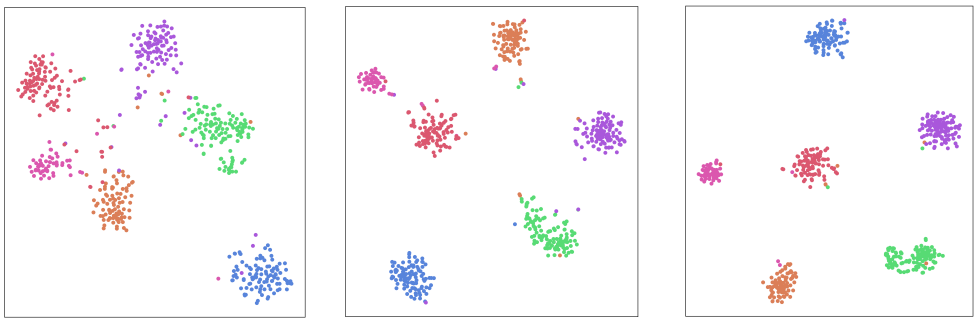
A IMPLEMENTATION DETAILS.

To ensure fair comparisons with previous work Liang et al. (2020a); Chen et al. (2022); Karim et al. (2023), we employ the ResNet-50 He et al. (2016) as the network backbone for the Office-Home and DomainNet-126 datasets, and ResNet-101 for the VisDA-C dataset. The network architecture follows the same configuration as SHOT Liang et al. (2020a). Specifically, we replace the original fully connected (FC) layer in ResNet-50/101 with a bottleneck layer of 256 dimensions and apply batch normalization Ioffe & Szegedy (2015). This modified setup serves as the feature extractor+projector head, producing feature representations and embedding of dimensions $d' = 2048$ and $d = 256$, respectively. Additionally, we include an extra fully connected layer with weight normalization Salimans & Kingma (2016) as a task-specific classifier.

In the first step of model pre-adaptation, we use a batch size of 64. The value of λ is set as $\lambda = \lambda_0 \cdot (1 + 10 \cdot p')^{-5}$, where $\lambda_0 = 1$, and p' represents the training progress variable ranging from 0 to 1, calculated as $\frac{iter}{max_iter}$. In the second step of hypothesis consolidation, we set the number of nearest/furthest neighbor per instance z as 3, and set hypothesis per instance $\tilde{k}$ as 4, respectively. The ranking thresholds τ_1 and τ_2 are determined as a percentage of the total number of samples on the three datasets, specifically set at 0.8% and 1.6%. In the third step of semi-supervised learning, we set the size of $\mathcal{B}_l$ and $\mathcal{B}_u$ to 64.

We use the SGD optimizer with a momentum of 0.9 and a weight decay of $1e^{-3}$ for all datasets. The learning rate is set as $1e^{-4}$ for all datasets, except for the bottleneck layer and the additional fully connected layer, where it is set as $1e^{-3}$. We train for 40 epochs on the Office-Home and DomainNet-126 datasets, where 9 epochs are dedicated to the model pre-adaptation. For the VisDA-C dataset, we train for 15 epochs, with 7 epochs allocated for the model pre-adaptation. All images from the datasets undergo augmentation, including weak and strong augmentation. Weak augmentation involves a standard flip-and-shift augmentation strategy, while strong augmentation is similar to the approach used in the work of Sohn et al. (2020).

B VISUALIZATION



(a) Source model only (b) PA only (c) Ours

Figure B1: t-SNE of target features on the randomly selected 6 DN-126 (Rw→Cl) classes.

In t-SNE visualization, we compare the results with the state before adaptation by examining three approaches: source model only, model pre-adaptation (PA), and our novel method. The source model only demonstrates shortcomings, experiencing false predictions within each class and struggling to establish clear intra-class boundaries. While PA generally achieves accurate predictions within each class, it falls short in generating clear intra-class boundaries. In contrast, our method excels in achieving accurate predictions within each class and successfully generates distinct intra-class boundaries, which showcases its ability to enhance prediction accuracy and produce well-defined intra-class boundaries.

Table 1: Comparison of pseudo-label quantity and quality on DomainNet-126 (Rw→Cl). Quantity (%) refers to the proportion of selected samples to total samples. Quality (%) refers to precision (%) of chosen sample.

DomainNet (Rw->Cl)	Quantity (%)	Quality (%)
source model only (con>0.95)	3.95	95.80
SHOT (con>0.95)	61.83	80.38
Ours (PA,con>0.95)	79.13	80.76
Ours (HCPR)	21.35	84.02
Ours (PA+HCPR)	24.65	90.76

C PSEUDO-LABEL QUANTITY AND QUALITY

In this section, we assess both the quality and quantity of pseudo-labels generated by each component of our method, comparing them with the source model alone and SHOT. Pseudo-label quantity is measured by the ratio of selected samples to the total samples, while pseudo-label quality is defined as the precision of the selected samples. The results are shown in Table 1. As seen, using the original source model generate good pseudo-label quality within the selected group, but only a small number of samples satisfy the high confidence condition. On the other hand, SHOT and PA select a large number of samples but with a relatively poor quality of approximately 80%. In comparison, our method achieves both good pseudo-label quality, approximately 90.76%, and a substantial quantity of pseudo-labels (24.65%). When comparing only HCPR and only PA, we observed that PA generates nearly four times as many pseudo-labels as HCPR but with lower quality. This suggests the presence of significant noise in the pseudo-labels generated by PA.

D THE COURSE OF TRAINING

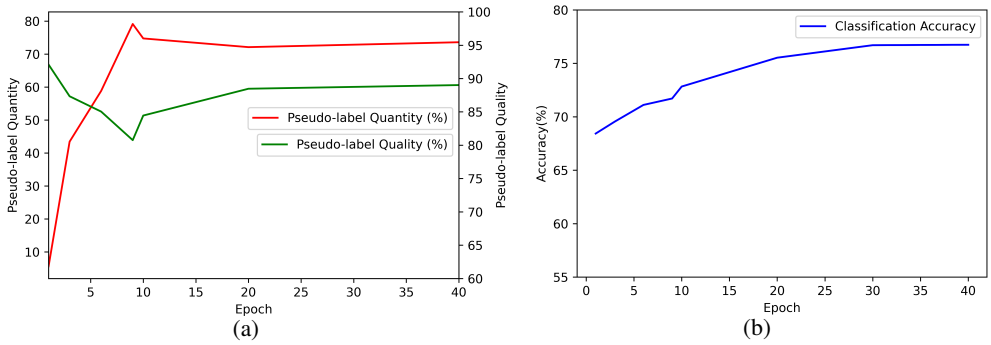


Figure D1: Pseudo-label quantity, quality and classification accuracy of our method over training on DomainNet-126 (Rw→Cl).

We conducted experiments to analyze the quantity, quality, and performance throughout the training process shown in Figure D1. Our findings revealed that in the initial step with PA (0-9th epochs), there is a significant increase in the quantity of pseudo labels, albeit accompanied by a gradual decrease in their quality. However, with the assistance of HCPR (after 9th epch, before 10th epoch), the quality of pseudo-labels experiences a significant increase, accompanied by a substantial quantity. In the subsequent third step involving FM (10-40th epochs), the quality of pseudo labels has a gradual improvement, which subsequently stabilizes at a consistent level.

E EFFECTIVENESS ANALYSIS.

We conducted an analysis and comparison of the memory usage and running time of our method with recent works, including AdaContrast Chen et al. (2022), C-SFDA Karim et al. (2023), and

Table E1: Effectiveness analysis on contrastive-based method and our methods. "BS" and "Mem" represent the batch size and peak memory on a single GPU. The running time is measured on 1 Tesla A100 GPU with 40 epochs.

DomainNet-126 (Rw→Cl)	Batch Size	Memory	Time	Accuracy
AdaConstrast Chen et al. (2022)	128	>32G	-	70.2
C-SFDA Karim et al. (2023)	256	>64G	-	70.8
GPL Litrico et al. (2023)	256	>64G	3h	74.2
Ours	128	17G	2h	**76.9**

GPL Litrico et al. (2023). Interestingly, our method requires normal memory usage, whereas the other methods consume more than 32GB of memory. Despite using standard memory, our approach achieves higher accuracy in comparison. Additionally, the running time of our method is considerably less than that of GPL.

F Further Exploration of Recursive HCPR Application

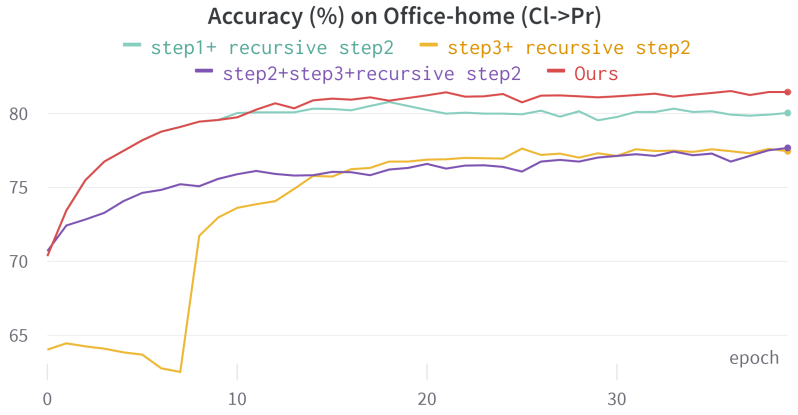


Figure F1: Accuracy of different components in our method with Recursive HCPR on the Office-Home (Cl→Pr) dataset.

We further investigate the influence of recursively applying HCPR to either PA or FixMatch. We conduct experiments using the Office-Home (Cl→Pr) dataset and configure the following scenarios:

- Combining Step 1 and Step 2, with Step 2 calculated at the 9th, 15th, and 20th epochs (indicated by the green curve in Figure F1).
- Combining Step 3 and Step 2, with Step 2 calculated at the 7th, 15th, and 20th epochs (indicated by the yellow curve in Figure F1).
- Combining Step 2 and Step 3, with Step 2 calculated at the 0th, 15th, and 20th epochs (indicated by the purple curve in Figure F1).
- Our method, represented by the red curve.

Our observations indicate that utilizing Step 2 only once is sufficient, and recursive HCPR application does not yield improvements. However, we do note that HCPR plays a crucial role in enhancing FixMatch, particularly in improving the quality of pseudo-labels.

G Impact of the z —the Number of Nearest and Furthest Neighbor

In the initial step of our model pre-adaptation, we select the z-nearest and z-furthest neighbors for each target sample. In this analysis, we examine the influence of the z value. Figure G1 and Table G1

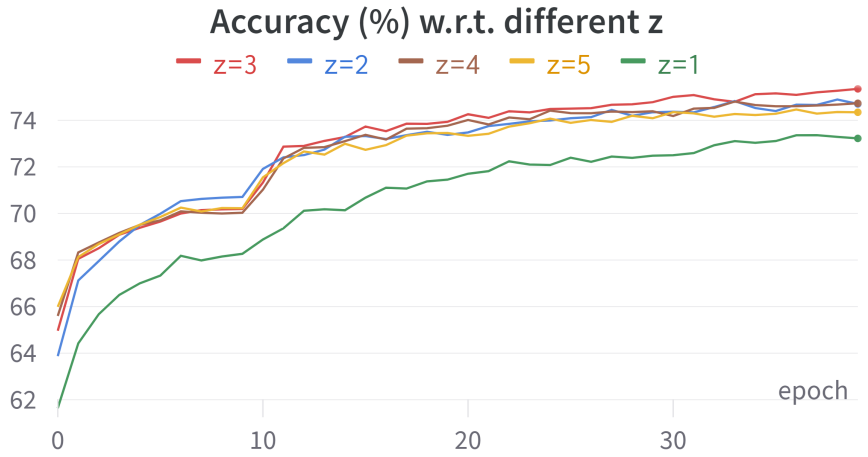


Figure G1: Accuracy of different z in our method on the **DomainNet-126** (Pt→Cl) dataset. When step 1 (0-9 epochs) achieves and maintains the best results, HCPR plays a pivotal role in enhancing the performance of the model.

Table G1: **DomainNet-126** (Pt→Cl) accuracy (%) of the proposed method with different number of the z.

z	1	2	3	4	5
Accuracy	73.2	74.7	**75.4**	74.7	74.3

showcase the performance throughout the training process and top-1 accuracy of classification on the DomainNet-126 (Pt→Cl) dataset for different values of z. The results indicate that even with just one nearest and furthest neighbor, we achieve favorable classification accuracy, and selecting 2-5 nearest and furthest neighbors yields optimal performance. Moreover, as observed in Figure G1, it is worth noting that when step 1 (0-9 epochs) achieves and maintains the best results, HCPR plays a pivotal role in enhancing the performance of the model.