Calibrating Without Labels: Source-Free Conformal Prediction Using Pseudo-Labels

Shachar Angelman SHACHAR.PARSHANI@BIU.AC.IL
Bar-Ilan University, Israel
Rotem Nizhar ROTEMNIZHAR10@GMAIL.COM
Tel-Aviv University, Israel
Jacob Goldberger JACOB.GOLDBERGER@BIU.AC.IL
Bar-Ilan University, Israel

Editor: Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

Abstract

We address the problem of conformal prediction (CP) in the challenging setting of source-free domain adaptation (SFDA), where models must be calibrated using only unlabeled data from the target domain. Existing CP methods for domain shift rely heavily on labeled source data and importance weighting (IW), but we demonstrate that these approaches perform poorly in practice, even when source labels are available. As an alternative, we propose Source-Free Conformal Prediction (SFCP), a simple and effective method that replaces the unavailable target labels with pseudo-labels generated by the source model. We show both theoretically and empirically that, despite their inherent noise, these pseudo-labels can be reliably used to estimate conformal thresholds. Our method requires no access to source data and no hyperparameter tuning, making it particularly suitable for real-world SFDA scenarios. Experiments across more than 100 domain shifts demonstrate that SFCP achieves coverage levels comparable to oracle CP while consistently outperforming IW-based methods.

Keywords: Conformal prediction, distribution shift, unsupervised domain shift, SFDA.

1. Introduction

In safety-critical applications of machine learning, such as medical diagnosis or autonomous driving, it is essential for models not only to predict accurately but also to express uncertainty in a reliable and interpretable manner. A powerful approach to uncertainty quantification is Conformal Prediction (CP) (Vovk et al., 2005), which constructs a prediction set that contains the true label with a user-specified confidence level, without assuming a specific data distribution. This capability makes CP especially valuable in high-stakes settings, where decisions must be both precise and trustworthy. CP generates a prediction set with a formal guarantee that the true class is included with at least the specified confidence level. The CP challenge is to minimize the size of the prediction set while upholding this confidence guarantee. As neural networks are increasingly applied to safety-critical fields, CP has become an essential calibration tool (Lu et al., 2022a,b; Olsson et al., 2022). Importantly, CP is a general framework rather than a specific algorithm. It involves constructing the prediction set using a nonconformity score with CP algorithms differing primarily in how this score is defined.

However, the deployment of deep neural networks in real-world environments frequently encounters domain shift—the phenomenon where the test data distribution differs from that of the training data. A possible solution is to adapt the network to the new domain. In the context of Unsupervised Domain Adaptation (UDA), it is assumed that unlabeled data from the target domain are available, but no annotations are provided. Numerous UDA approaches have been developed, including adversarial training techniques that align the source and target domain distributions (Ganin et al., 2016) and self-training algorithms that generate pseudo-labels for the target domain data (Zou et al., 2019).

In a typical UDA setup, adaptation is performed using labeled source domain data alongside unlabeled target domain data. In practice, access to source data is restricted due to privacy, legal, or proprietary constraints. This is common in fields such as healthcare or finance. In a Source-Free Domain Adaptation (SFDA) setup access to source domain data during the adaptation process is prohibited. Consequently, SFDA relies on unsupervised learning and self-training techniques. Most SFDA methods primarily focus on self-training using pseudo-labels for the target domain and entropy minimization strategies. However, due to domain shifts, these pseudo-labels are often very noisy. To mitigate this issue, various approaches have been proposed to refine pseudo-labels during training (e.g., (Chen et al., 2022; Karim et al., 2023; Yi et al., 2023)). A common strategy involves updating pseudo-labels iteratively at each epoch to improve their alignment with the target domain distribution (Liang et al., 2020). Zhang et al. (2023) addressed the noise problem by leveraging a robust pre-trained network to generate pseudo-labels while filtering out low-confidence samples. Diamant et al. (2024) applied a noise-robust training technique to handle the label noise in the pseudo-labels. While CP has seen growing interest in domain adaptation settings, its application to the challenging SFDA setting — where neither target labels nor source data are available — remains unexplored.

Here, we address the problem of applying a CP procedure to the model adapted to the target domain obtained by either a UDA or an SFDA process where in both cases the data available from the target domain are unlabeled. The most common approach that applies CP to unlabeled data from a new domain is based on Importance Weighting (IW), which assigns higher weights to source examples that resemble those in the target domain. Tibshirani et al. (2019) studied CP under covariate shifts, where the distribution of the conditional label distribution remains unchanged. In the UDA setup where labeled data from the source domain is available, we can apply IW on the labeled data from the source domain, but accurately estimating the exact likelihood-ratio weights is challenging in high-dimensional data such as images. In SFDA setups where no data from the source domain is available, IW-based CP algorithms are not relevant.

IW and its many variants are currently the standard approach for CP in covariate shift setups (Wang et al., 2020; Pampari and Ermon, 2020). We evaluated the applicability of IW to UDA setups using standard publicly available domain shift image classification datasets. To the best of our knowledge, this is the first exhaustive evaluation of IW in a UDA setup. Here, we address CP in the context of networks trained on a source domain and then adapted to a target domain in an unsupervised manner. We are also unaware of any such evaluation in the simpler case of calibrating the source model on a target domain without labeled data from the new domain. Our results show that, in practice, IW-based methods do not perform well, and we analyze the reasons why.

In this paper, we propose a new approach to CP under domain shift that is practical, effective, and fully source-free. Our method, called Source-Free Conformal Prediction (SFCP), leverages pseudo-labels generated by a source model to calibrate the adapted network on the target domain. While pseudo-labels are inherently noisy, we demonstrate both theoretically and empirically that they can yield reliable CP thresholds. We introduce a probabilistic model that captures the structure of pseudo-label noise and provides formal coverage guarantees. Through extensive experiments we show that SFCP consistently outperforms IW-based methods, despite using no labeled data at all.

Our contributions are the following. We empirically demonstrate that IW-based CP performs poorly under domain shift, even with access to source labels. We propose a simple and practical Source-Free Conformal Prediction (SFCP) method that uses pseudo-labels without requiring source data. We provide a theoretical justification for using pseudo-labels in CP calibration, introducing a generative model under which coverage guarantees hold. We validate SFCP on over 100 source-target pairs across 5 datasets, showing consistently strong coverage and efficiency compared to IW-based methods.

2. Background and related works

Consider a setup involving a classification network that categorizes an input x into K predetermined classes. Given a coverage level of $1-\alpha$, we aim to identify the smallest possible prediction set (a subset of these classes) ensuring that the correct class is within the set with a probability of at least $1-\alpha$. A straightforward strategy to achieve this objective involves sequentially incorporating classes from the highest to the lowest probabilities until their cumulative sum exceeds the threshold of $1-\alpha$. Despite the network's output adopting a mathematical distribution format, it does not inherently reflect the actual class distribution. Typically, the network is not calibrated and it tends to be overly optimistic (Guo et al., 2017). Consequently, this straightforward approach does not ensure the inclusion of the correct class with the desired probability.

The first step of the CP algorithm involves forming a nonconformity score $S(x,y)$ that measures the network's uncertainty between x and its true label y (larger scores indicate worse agreement). The Homogeneous Prediction Sets (HPS) score (Vovk et al., 2005) is $S_{\text{HPS}}(x,y) = 1 - p(y|x;\theta)$, s.t. θ is the network parameter set. The Adaptive Prediction Score (APS) (Romano et al., 2020) is the sum of all class probabilities that are not lower than the probability of the candidate class y:

$$S_{APS}(x,y) = \sum_{\{i|p_i \ge p_y\}} p_i, \tag{1}$$

such that $p_i = p(y=i|x;\theta)$ and p_y is the probability of the label y. The RAPS score (Angelopoulos et al., 2021) is a variant of APS, which is defined as follows:

$$S_{RAPS}(x,y) = \sum_{\{i|p_i \ge p_y\}} p_i + a \cdot \max(0, (NC - b)) \tag{2}$$

s.t. $NC = |\{i|p_i \ge p_y\}|$ and a, b are parameters that need to be tuned. RAPS is especially effective in the case of a large number of classes where it explicitly encourages small prediction sets.

We can also define a randomized version of the nonconformity score. For example in the case of APS we define:

$$S_{APS}(x, y, u) = \sum_{\{i | p_i > p_y\}} p_i + u \cdot p_y, \quad u \sim U[0, 1]. \tag{3}$$

The random version tends to yield the required coverage more precisely and thus produces smaller prediction sets (Angelopoulos et al., 2023). The CP prediction set of a data point x is defined as $C_q(x) = \{y | S(x, y) \le q\}$ where q is a threshold found using a labeled validation set $(x_1, y_1), ..., (x_n, y_n)$. The CP theorem states that if we set q to be the $(1-\alpha)$ quantile of the conformal scores $S(x_1, y_1), ..., S(x_n, y_n)$ we can guarantee that $1-\alpha \le p(y \in C_q(x)) \le 1-\alpha + \frac{1}{n+1}$, where x is a test point and y is its the true label (Vovk et al., 2005). In the random case there is still a coverage guarantee, which is defined by marginalizing over all test points x and samplings u from the uniform distribution (Romano et al., 2020). Note that the coverage guarantee is for a marginal probability over all possible test points and coverage may be worse or better for different points. It can be proved that obtaining a conditional coverage guarantee is impossible (Foygel Barber et al., 2021).

In recent years, there have been many attempts to address the problem of conformal prediction under distribution shift. Tibshirani et al. (2019) adopted importance weighting by the likelihood ratio and coverage is ensured under covariate shift. Barber et al. (2023) generalized the reweighting idea to handle more general distribution shifts, but choosing the weights is generally unclear in practice. Barber et al. (2023), Angelopoulos et al. (2022), and Angelopoulos et al. (2023) upper-bound coverage gap via total variation distance. Gibbs and Candes (2021), Xu and Xie (2021), and Gibbs and Candès (2024) focus on CP under dynamic shift (test distribution changes over time). Works that focus on static domain shift either modify vanilla CP upon a residual-driven model for robust coverage (Gendler et al., 2021; Roth, 2022; Cauchois et al., 2024; Zou and Liu, 2024) or incorporate a conformal-based loss during training to obtain robust and efficient prediction sets (Yan et al., 2024). These methods are sensitive to the selected hyperparameters, and the obtained prediction sets are overly conservative. Other approaches consider CP under multi-source domain generalization, which focuses on developing a model that generalizes effectively to a new test distribution (Sagawa et al., 2019; Krueger et al., 2021). A related topic is federated CP, which aims to train a model across decentralized data sources to perform well on a known test distribution (typically a uniformly weighted mixture of source distributions) without requiring centralization to ensure privacy (Lu et al., 2023; Humbert et al., 2023; Plassier et al., 2023). Other recent works include (Ai and Ren, 2024; Ge et al., 2024). All these research directions and studies, however, cannot be applied to SFDA scenarios where access to the source domain is restricted.

3. Source-Free Conformal Prediction

Problem statement. We first formulate the problem of CP in the UDA and SFDA setups. Assume that a network originally trained on the source domain was adapted to the target domain in an unsupervised manner, either with or without access to the labeled data from the source domain. We are given an unlabeled target domain dataset $\mathcal{T} = \{x_t^i\}_{i=1}^{n_t}$ with n_t samples. In the SFDA this is all we have and in the simpler UDA case, we are also given a labeled source domain validation-set dataset, denoted as $\mathcal{S} = \{(x_s^i, y_s^i)\}_{i=1}^{n_s}$ with n_s samples. Here, our goal is to apply CP on the predictions of the adapted network to find a threshold q that satisfies $p(y \in C_q(x)) \geq 1 - \alpha$ where (x, y) are sampled from the target domain.

Importance Weighting. We start by briefly reviewing the IW method (Tibshirani et al., 2019). For each sample x from the source domain, define the likelihood ratio $w(x) = f(x|C = 1)/f(x|C = 0)$ such that C is a binary random variable, and the values 0 and 1 stand for the source and target domain, respectively. We then apply a weighted variant of the CP procedure to the labeled data from the source domain. We expect that source domain samples that are more similar to those from the target domain are more relevant when calibrating the network on the target domain. A covariate shift occurs where the image distribution changes between the source and target domains but the conditional distribution of the class given the image remains the same. In the case of covariate shift the IW method provides a theoretical coverage guarantee (Tibshirani et al., 2019; Barber et al., 2023).

In practice, the domain-dependent sample distribution $f(x|C)$ is unknown and we need to estimate the weights of the source samples from the data. Direct estimation of the likelihood ratio for high-dimensional data, such as images, is very challenging (see e.g. (Cauchois et al., 2024)). Tibshirani et al. (2019) proposed a heuristic replacement for the likelihood ratio which is based on using the given samples from the source and target domains to train a binary domain classifier $p(C|x)$. Then, we can approximate the likelihood-ratio weight by $\tilde{w}(x) = p(C = 1|x)/p(C = 0|x)$ such that we assign higher weights to images from the source domain that are classified as target domain data.

We evaluated the effectiveness of the approximated IW method on several standard domain-shift datasets containing over one hundred pairs of source and target domains. The experimental results (see Section 4) demonstrate that the IW method performs poorly on these datasets. Here, we analyze the reasons for this observation. The main justification of IW is that source samples that are classified as targets are more relevant when calibrating the adapted network on the target domain. Therefore, applying a CP to these samples results in a CP threshold that is more similar to the oracle threshold obtained by using labeled data from the target domain. To verify this assumption, we compared IW to two alternatives. The first used the opposite weights $p(C = 0|x)/p(C = 1|x)$, and the second was an unweighted version that calibrated the adapted model using the labeled data from the source domain. We also implemented the oracle calibration using labeled samples from the target domain. Fig. 1 illustrates the CP thresholds obtained for all the source-target domain pairs from several standard domain-shift datasets (see details in Section 4). The results showed that there is no correlation between the weight $p(C = 1|x)$ and the relevance of a source domain sample x to calibrating the adapted model on the target domain. Thus, the approximated weights used in the IW method are ineffective and not directly related to the likelihood-ratio weights. To summarize, the IW approach with theoretical coverage guarantees is not applicable because we cannot compute the sample weights and also the covariate shift assumption is not necessarily satisfied for the evaluated domain shift datasets. The approximated procedure, which is based on training a binary domain classifier, has no theoretical coverage guarantee and produces poor results. Another drawback of IW (precise or approximated) is that it cannot be used if access to the source data is restricted.

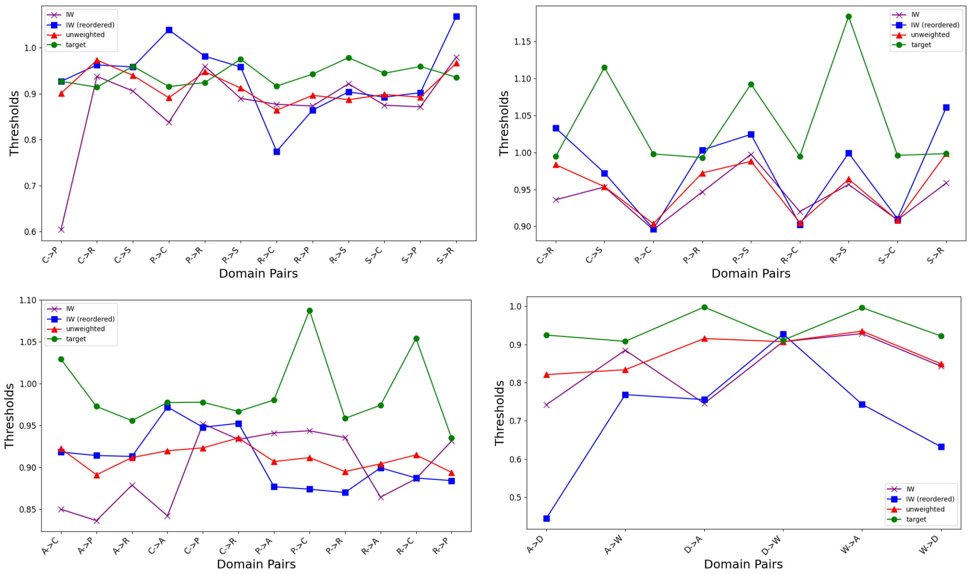

Figure 1: Comparison of thresholds for different conformal prediction strategies across all domain pairs. The strategies include importance weighting (IW) with original weights, reordered weights, and unweighted, using source data. Additionally, the target data is used for comparison. Results are shown for the following datasets: DomainNet-40, DomainNet-126, Office-Home, and Office-31, using the SFDA method DCPL (Diamant et al., 2024).

Pseudo Labels. Next, we present a simple domain shift CP method based on pseudo-labels, that produces good CP results on the target domain. This method does not require access to labeled data from the source domain and can, therefore, be used in an SFDA setup, where IW-based methods are not applicable. Pseudo-labeling is the most successful method for source-free domain adaptation; here we show that pseudo-labels are also very effective for CP under domain shift. Pseudo-labels are generated by applying the source model to the target domain data. Following a standard practice in SFDA methods, to reduce the pseudo-labels' noise level we can use unsupervised techniques and self-training (Zhang et al., 2023; Liang et al., 2020)). This involves utilizing the source model's predictions on target data along with a pre-trained strong feature extractor f_p (Swin-B) (Liu et al., 2021), to create centroids for each class. Cosine distance is then used to assign each example to its nearest centroid. This procedure reduces the noise level of the pseudo-labels. However, these pseudo-labels remain noisy and do not align perfectly with the true labels. (In our experiments on more than 100 domain pairs, this procedure reduced the average labeling error from 40% to 20%). In source-free domain adaptation, we cannot directly use pseudo-labels (PL) as they are highly noisy, requiring explicit handling of the label noise problem (Diamant et al., 2024). Our key novel claim is that, despite their high noise level, pseudo-labels can be directly utilized to find the CP threshold. Our observation is that PL noise follows a specific pattern such that, despite its high noise level, the estimated CP threshold

Algorithm 1 Source-Free Conformal Prediction (SFCP)

Input: A model trained on the source domain and unlabeled data from the target domain divided into training and calibration subsets.

- Compute pseudo-labels for the target samples using the source model. First, calculate class centroids as a weighted average of the features $f_p(x)$:

  $$C_k = \frac{\sum_i p(y_i = k|x_i) f_p(x_i)}{\sum_i p(y_i = k|x_i)}, \qquad k = 1, ..., K$$

  where $p(y_i = k|x_i)$ is the class probability of the target sample x_i based on the source model. For each target sample x_i, generate a pseudo-label $\tilde{y}_i$ based on its nearest centroid using the cosine distance:

  $$\tilde{y}_i = \arg\min_k \cos(C_k, f_p(x_i))$$

- Source-free domain adaptation: Apply any SFDA procedure using the unlabeled data from the target domain to adapt the source model to the target domain.

- Source-free domain calibration: Apply a standard CP procedure to calibrate the adapted model on the target domain such that pseudo-labels are used to replace the unavailable true labels.

of the adapted model on the target domain — whether calculated using true labels or pseudo-labels — yields similar results. Intuitively, we expect to observe incorrect pseudo-labels in cases where the adapted network struggles between two options. In that case both the true label y and pseudo-label $\tilde{y}$ seem plausible for the adapted classifier so the distributions of the scores $S(x, y)$ and $S(x, \tilde{y})$ are close to each other.

Next, we propose a probabilistic model for pseudo labels and derive a theoretical coverage guarantee based on that model. The problem of pseudo-labels is that they are a noisy version of the correct labels. Several studies (e.g. (Sesia et al., 2024; Clarkson et al., 2024)) analyzed the problem of applying CP to validation sets with noisy labels by assuming a class-dependent noise, i.e. $p(\tilde{y}|y, x) = p(\tilde{y}|y)$ where x is a sample from the target domain and y and $\tilde{y}$ are the corresponding true label and pseudo-label respectively. This noise model is relevant in various differential privacy schemes where noise is explicitly injected into the labels. However, this model does not apply to pseudo-labels, which are more likely to be incorrect in situations where the correct class is ambiguous given the image. We propose the following model for pseudo labels. Let $p(y|x)$ be the (true but unknown) conditional distribution of the correct label y given the image x. We assume that $\tilde{y}$ is independently sampled from the same conditional distribution, i.e.

$$p(x, y, \tilde{y}) = p(x) p_{y|x}(y|x) p_{y|x}(\tilde{y}|x). \tag{4}$$

The conditional probability that the pseudo-label is correct is thus $\sum_i p(y = i|x)^2$. In this model, for higher values of the probability of the correct class where the image class is easily determined, it is more likely that $\tilde{y} = y$ since it was sampled from the same conditional distribution. In other cases where the correct class is not clearly determined from the image, the pseudo-label is more likely to be wrong. For each scalar q define: $F(q) = p(y \in C_q(x)) = p(S(x,y) \leq q)$. $F(q)$ is thus the Cumulative Distribution Function (CDF) of the conformal score function $S(x,y)$. Similarly denote $\tilde{F}(q) = p(\tilde{y} \in C_q(x))$ where $\tilde{y}$ is the corresponding pseudo-label. Eq. (4) implies that $F(q) = \tilde{F}(q)$ for every q. We next prove that, assuming the pseudo-label modeling (4) we can use the pseudo-labels instead of the correct labels and obtain the same coverage guarantee.

Theorem 1 *Let q be the $1-\alpha$ threshold obtained by applying CP on target samples with pseudo-labels $(x_1, \tilde{y}_1), ..., (x_n, \tilde{y}_n)$. The assumption (4) implies that $p(y \in C_q(x)) \geq 1-\alpha$.*

Proof The CP theorem (Vovk et al., 2005) implies that $p(S(x,\tilde{y}) \leq q) = p(\tilde{y} \in C_q(x)) \geq 1-\alpha$. The pseudo-label model assumption (4) implies that $p(S(x,y) \leq q) = p(S(x,\tilde{y}) \leq q)$ which finally implies that $p(y \in C_q(x)) = p(\tilde{y} \in C_q(x)) \geq 1-\alpha$. ■

We note that the proof remains valid under a much weaker assumption than (4). We only need that the r.v. $S(x,y)$ and $S(x,\tilde{y})$ have the same distribution. In practice, we expect the two distributions F and $\tilde{F}$ to be similar, but not exactly the same. If we can upper bound the Kolmogorov-Smirnov distance $\text{KS}(F,\tilde{F}) = \max_q |F(q) - \tilde{F}(q)|$, we can state the following coverage guarantee:

Theorem 2 *Let q be the $1-\alpha$ threshold obtained by applying CP on target samples with pseudo-labels $(x_1, \tilde{y}_1), ..., (x_n, \tilde{y}_n)$. Denote $\Delta = KS(F,\tilde{F})$ then $p(y \in C_q(x)) \geq 1-\alpha-\Delta$.*

Proof The CP coverage guarantee theorem directly implies that $\tilde{F}(q) = p(\tilde{y} \in C_q(x)) \geq 1-\alpha$. Note that $\tilde{F}(q) = p(\tilde{y} \in C_q(x))$ is the marginal distribution over n validation samples and a test sample from the joint images and pseudo-labels distribution. Combining it with the assumption that $\text{KS}(F,\tilde{F}) < \Delta$, we obtain $F(q) \geq \tilde{F}(q) - \text{KS}(F,\tilde{F}) \geq 1-\alpha-\Delta$ which yields the desired coverage guarantee. ■

We thus propose directly using the pseudo-labels to calibrate the adapted model on the target domain in spite of their high noise level. We dub this CP algorithm Source-Free Conformal Prediction (SFCP). The SFCP algorithm is summarized in Algorithm Box 1. In this section we proposed a probabilistic model for the pseudo-labels' noise. The main question is whether the model aligns with real data behavior. In the next section we empirically validate that this is indeed the case. Based on a statistics of $\Delta = \text{KS}(F,\tilde{F})$ across more than 100 source-target domain pairs, we show that indeed Δ is small and thus the coverage guarantee is informative. More than that we show that the finite-set estimations of $F(q)$ and $\tilde{F}(q)$ pass the Kolmogorov–Smirnov (KS) test, i.e., there is no statistically significant evidence to reject the hypothesis that the data samples for $S(x,y)$ and from $S(x,\tilde{y})$ come from the same distribution.

4. Experiments

In this section, we evaluate the capabilities of our SFCP technique and the current IW-CP technique on a target domain after applying a SFDA procedure.

Table 1: RAPS calibration results for $1-\alpha = 0.9$ on **DomainNet40**, across various SFDA classification tasks and methods with different CP methods. We report the mean and the std over 1000 different splits.

SFDA	Method	C→P		C→R		C→S		P→C	
		size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)
DCPL	CP (oracle)	3.08 ± 0.17	0.90 ± 0.01	1.41 ± 0.05	0.90 ± 0.01	4.03 ± 0.22	0.90 ± 0.02	4.01 ± 0.28	0.90 ± 0.02
	SFCP	3.03 ± 0.16	0.90 ± 0.01	1.40 ± 0.05	0.90 ± 0.01	4.44 ± 0.23	0.91 ± 0.01	4.27 ± 0.30	0.91 ± 0.02
	IW-CP	1.60 ± 0.64	0.73 ± 0.09	1.27 ± 0.31	0.85 ± 0.06	1.80 ± 0.82	0.71 ± 0.10	1.73 ± 0.80	0.67 ± 0.12
SHOT	CP (oracle)	4.16 ± 0.21	0.90 ± 0.01	1.98 ± 0.08	0.90 ± 0.01	5.05 ± 0.26	0.90 ± 0.02	5.20 ± 0.34	0.90 ± 0.02
	SFCP	3.99 ± 0.21	0.89 ± 0.01	1.94 ± 0.07	0.90 ± 0.01	5.54 ± 0.27	0.91 ± 0.01	5.26 ± 0.31	0.90 ± 0.02
	IW-CP	2.73 ± 1.22	0.79 ± 0.10	2.16 ± 0.41	0.90 ± 0.03	2.80 ± 1.14	0.76 ± 0.09	2.00 ± 0.96	0.60 ± 0.15
AaD	CP (oracle)	3.59 ± 0.18	0.90 ± 0.01	1.53 ± 0.05	0.90 ± 0.01	4.38 ± 0.22	0.90 ± 0.02	4.51 ± 0.30	0.90 ± 0.02
	SFCP	4.19 ± 0.20	0.93 ± 0.01	1.51 ± 0.05	0.90 ± 0.01	5.99 ± 0.27	0.94 ± 0.01	5.47 ± 0.31	0.93 ± 0.01
	IW-CP	2.47 ± 1.04	0.80 ± 0.09	1.54 ± 0.24	0.89 ± 0.03	2.50 ± 1.19	0.75 ± 0.11	1.63 ± 1.15	0.56 ± 0.19

SFDA	Method	P→R		P→S		R→C		R→P	
		size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)
DCPL	CP (oracle)	1.47 ± 0.05	0.90 ± 0.01	4.87 ± 0.27	0.90 ± 0.02	3.90 ± 0.28	0.90 ± 0.02	3.04 ± 0.18	0.90 ± 0.01
	SFCP	1.36 ± 0.05	0.88 ± 0.01	5.08 ± 0.26	0.90 ± 0.01	3.96 ± 0.28	0.90 ± 0.02	2.82 ± 0.16	0.89 ± 0.01
	IW-CP	2.84 ± 0.18	0.93 ± 0.02	2.31 ± 0.91	0.75 ± 0.08	3.45 ± 0.76	0.86 ± 0.04	5.75 ± 0.40	0.90 ± 0.02
SHOT	CP (oracle)	2.13 ± 0.08	0.90 ± 0.01	5.31 ± 0.27	0.90 ± 0.01	4.96 ± 0.34	0.90 ± 0.02	4.15 ± 0.22	0.90 ± 0.01
	SFCP	1.95 ± 0.08	0.88 ± 0.01	6.25 ± 0.30	0.92 ± 0.01	5.04 ± 0.33	0.90 ± 0.02	3.96 ± 0.21	0.89 ± 0.01
	IW-CP	1.86 ± 0.40	0.87 ± 0.04	3.21 ± 1.14	0.78 ± 0.07	4.38 ± 1.41	0.86 ± 0.05	3.73 ± 0.49	0.88 ± 0.02
AaD	CP (oracle)	1.48 ± 0.05	0.90 ± 0.01	4.58 ± 0.23	0.90 ± 0.02	4.50 ± 0.31	0.90 ± 0.02	3.56 ± 0.19	0.90 ± 0.01
	SFCP	1.45 ± 0.05	0.90 ± 0.01	6.66 ± 0.27	0.94 ± 0.01	5.74 ± 0.35	0.93 ± 0.01	4.15 ± 0.21	0.92 ± 0.01
	IW-CP	1.42 ± 0.20	0.88 ± 0.03	3.05 ± 0.93	0.77 ± 0.08	4.41 ± 1.00	0.89 ± 0.03	2.97 ± 0.45	0.87 ± 0.03

SFDA	Method	R→S		S→C		S→P		S→R	
		size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)
DCPL	CP (oracle)	5.21 ± 0.29	0.90 ± 0.02	3.90 ± 0.28	0.90 ± 0.02	3.38 ± 0.18	0.90 ± 0.01	1.61 ± 0.06	0.90 ± 0.01
	SFCP	4.66 ± 0.25	0.89 ± 0.01	3.70 ± 0.25	0.89 ± 0.02	3.10 ± 0.16	0.89 ± 0.01	1.43 ± 0.05	0.88 ± 0.01
	IW-CP	3.27 ± 0.62	0.81 ± 0.04	1.92 ± 1.16	0.69 ± 0.14	1.00 ± 0.40	0.57 ± 0.13	2.89 ± 0.31	0.88 ± 0.04
SHOT	CP (oracle)	6.00 ± 0.31	0.90 ± 0.02	4.84 ± 0.34	0.90 ± 0.02	4.07 ± 0.21	0.90 ± 0.01	2.14 ± 0.08	0.90 ± 0.01
	SFCP	6.32 ± 0.32	0.91 ± 0.01	4.81 ± 0.31	0.90 ± 0.02	3.92 ± 0.20	0.90 ± 0.01	1.87 ± 0.07	0.88 ± 0.01
	IW-CP	3.38 ± 1.09	0.79 ± 0.06	2.08 ± 1.15	0.66 ± 0.14	1.21 ± 0.68	0.56 ± 0.13	3.49 ± 0.38	0.89 ± 0.04
AaD	CP (oracle)	5.08 ± 0.26	0.90 ± 0.01	4.12 ± 0.28	0.90 ± 0.02	3.49 ± 0.18	0.90 ± 0.01	1.50 ± 0.05	0.90 ± 0.01
	SFCP	7.78 ± 0.31	0.94 ± 0.01	5.31 ± 0.33	0.94 ± 0.01	4.09 ± 0.20	0.92 ± 0.01	1.60 ± 0.05	0.91 ± 0.01
	IW-CP	2.41 ± 0.90	0.73 ± 0.09	1.50 ± 1.03	0.55 ± 0.20	1.09 ± 0.55	0.52 ± 0.15	2.84 ± 0.17	0.90 ± 0.03

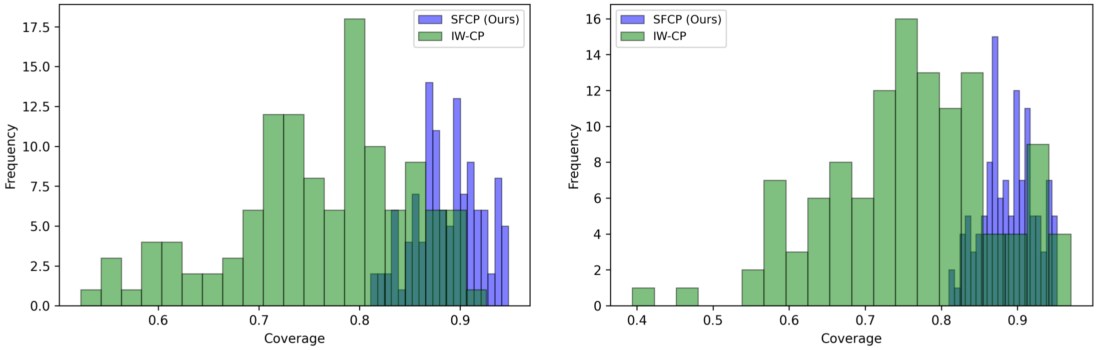

Figure 2: Histogram of coverage for SFCP and IW-CP across all 102 source-target pairs used in our experiments, using rand-RAPS (left) and HPS (right) and $1-\alpha = 0.9$.

Table 2: RAPS calibration results for $1{-}\alpha = 0.9$ on **DomainNet126**, across various SFDA classification tasks and methods with different CP methods. We report the mean and the std over 1000 different splits.

SFDA	Method	C→R		C→S		P→C	
		size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)
DCPL	CP (oracle)	3.08 ± 0.09	0.90 ± 0.01	19.66 ± 0.68	0.90 ± 0.01	8.43 ± 0.33	0.90 ± 0.01
	SFCP	1.92 ± 0.05	0.85 ± 0.01	12.14 ± 0.54	0.88 ± 0.01	6.90 ± 0.23	0.88 ± 0.01
	IW-CP	1.52 ± 0.16	0.80 ± 0.02	3.14 ± 0.37	0.71 ± 0.02	2.65 ± 0.58	0.70 ± 0.04
SHOT	CP (oracle)	4.41 ± 0.11	0.90 ± 0.01	14.61 ± 0.60	0.90 ± 0.01	10.68 ± 0.34	0.90 ± 0.01
	SFCP	3.09 ± 0.09	0.86 ± 0.01	12.48 ± 0.52	0.89 ± 0.01	9.35 ± 0.31	0.88 ± 0.01
	IW-CP	2.08 ± 0.29	0.78 ± 0.03	4.55 ± 0.58	0.73 ± 0.03	4.50 ± 1.35	0.72 ± 0.06
AaD	CP (oracle)	6.42 ± 0.15	0.90 ± 0.01	19.08 ± 0.72	0.90 ± 0.01	12.46 ± 0.38	0.90 ± 0.01
	SFCP	4.27 ± 0.10	0.87 ± 0.01	36.63 ± 0.93	0.94 ± 0.01	21.98 ± 0.81	0.94 ± 0.01
	IW-CP	1.56 ± 0.13	0.75 ± 0.02	4.79 ± 0.32	0.74 ± 0.02	4.26 ± 0.69	0.71 ± 0.04

SFDA	Method	P→R		P→S		R→C	
		size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)
DCPL	CP (oracle)	2.65 ± 0.07	0.90 ± 0.01	16.13 ± 0.68	0.90 ± 0.01	6.92 ± 0.28	0.90 ± 0.01
	SFCP	1.78 ± 0.05	0.85 ± 0.01	21.66 ± 0.69	0.91 ± 0.01	6.76 ± 0.29	0.90 ± 0.01
	IW-CP	1.58 ± 0.11	0.83 ± 0.01	17.84 ± 0.26	0.88 ± 0.01	3.28 ± 0.21	0.79 ± 0.02
SHOT	CP (oracle)	3.66 ± 0.09	0.90 ± 0.01	15.29 ± 0.69	0.90 ± 0.01	9.48 ± 0.35	0.90 ± 0.01
	SFCP	2.64 ± 0.07	0.86 ± 0.01	15.49 ± 0.70	0.90 ± 0.01	9.55 ± 0.36	0.90 ± 0.01
	IW-CP	1.95 ± 0.17	0.80 ± 0.02	9.92 ± 0.36	0.86 ± 0.01	4.79 ± 0.28	0.78 ± 0.02
AaD	CP (oracle)	3.32 ± 0.10	0.90 ± 0.01	18.55 ± 0.66	0.90 ± 0.01	8.87 ± 0.32	0.90 ± 0.01
	SFCP	2.32 ± 0.06	0.87 ± 0.01	42.37 ± 1.04	0.95 ± 0.01	23.39 ± 0.69	0.95 ± 0.01
	IW-CP	1.55 ± 0.10	0.80 ± 0.02	7.25 ± 0.29	0.80 ± 0.01	3.99 ± 0.24	0.78 ± 0.02

SFDA	Method	R→S		S→C		S→R	
		size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)
DCPL	CP (oracle)	24.99 ± 1.14	0.90 ± 0.01	6.31 ± 0.28	0.90 ± 0.01	3.33 ± 0.10	0.90 ± 0.01
	SFCP	13.42 ± 0.55	0.87 ± 0.01	5.68 ± 0.23	0.89 ± 0.01	1.88 ± 0.05	0.85 ± 0.01
	IW-CP	3.93 ± 0.20	0.73 ± 0.01	3.10 ± 0.28	0.80 ± 0.02	1.68 ± 0.12	0.83 ± 0.01
SHOT	CP (oracle)	21.02 ± 1.21	0.90 ± 0.01	7.86 ± 0.32	0.90 ± 0.01	4.39 ± 0.10	0.90 ± 0.01
	SFCP	19.48 ± 0.88	0.90 ± 0.01	7.81 ± 0.31	0.90 ± 0.01	3.08 ± 0.08	0.86 ± 0.01
	IW-CP	5.39 ± 0.29	0.73 ± 0.02	4.29 ± 0.38	0.80 ± 0.02	2.47 ± 0.19	0.82 ± 0.01
AaD	CP (oracle)	25.20 ± 0.82	0.90 ± 0.01	7.04 ± 0.30	0.90 ± 0.01	5.65 ± 0.13	0.90 ± 0.01
	SFCP	47.41 ± 1.07	0.94 ± 0.01	15.41 ± 0.65	0.95 ± 0.01	3.51 ± 0.10	0.87 ± 0.01
	IW-CP	5.44 ± 0.23	0.74 ± 0.01	3.93 ± 0.51	0.80 ± 0.03	5.01 ± 0.19	0.86 ± 0.01

Table 3: RAPS calibration results for $1-\alpha = 0.9$ on **Office-31**, across various SFDA classification tasks and methods with different CP methods. We report the mean and the std over 1000 different splits.

SFDA	Method	A→D		A→W		D→A	
		size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)
DCPL	CP (oracle)	1.37 ± 0.40	0.92 ± 0.07	1.19 ± 0.24	0.91 ± 0.06	4.99 ± 0.74	0.90 ± 0.03
	SFCP	1.36 ± 0.38	0.92 ± 0.07	1.43 ± 0.32	0.94 ± 0.04	2.18 ± 0.26	0.84 ± 0.03
	IW-CP	0.94 ± 0.55	0.71 ± 0.21	0.90 ± 0.16	0.80 ± 0.09	1.14 ± 0.52	0.63 ± 0.14
SHOT	CP (oracle)	2.00 ± 0.62	0.91 ± 0.07	1.77 ± 0.44	0.91 ± 0.06	6.19 ± 0.72	0.90 ± 0.03
	SFCP	1.95 ± 0.60	0.91 ± 0.07	1.92 ± 0.48	0.92 ± 0.05	2.94 ± 0.34	0.84 ± 0.04
	IW-CP	1.29 ± 1.08	0.69 ± 0.24	1.00 ± 0.33	0.74 ± 0.12	1.33 ± 0.73	0.59 ± 0.15
AaD	CP (oracle)	1.75 ± 0.52	0.91 ± 0.07	1.32 ± 0.27	0.91 ± 0.05	5.87 ± 0.72	0.90 ± 0.03
	SFCP	1.87 ± 0.54	0.93 ± 0.06	1.77 ± 0.46	0.94 ± 0.04	2.54 ± 0.28	0.84 ± 0.04
	IW-CP	1.14 ± 0.79	0.71 ± 0.21	0.92 ± 0.32	0.75 ± 0.12	1.18 ± 0.55	0.62 ± 0.14
SFDA	Method	D→W		W→A		W→D	
		size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)
DCPL	CP (oracle)	1.08 ± 0.18	0.91 ± 0.05	5.45 ± 0.71	0.90 ± 0.03	1.17 ± 0.29	0.92 ± 0.07
	SFCP	1.05 ± 0.17	0.91 ± 0.06	1.96 ± 0.24	0.82 ± 0.04	1.16 ± 0.28	0.91 ± 0.07
	IW-CP	1.02 ± 0.47	0.83 ± 0.12	1.62 ± 0.44	0.72 ± 0.09	0.97 ± 0.36	0.81 ± 0.13
SHOT	CP (oracle)	1.29 ± 0.28	0.91 ± 0.05	6.62 ± 0.90	0.90 ± 0.03	1.37 ± 0.39	0.92 ± 0.07
	SFCP	1.24 ± 0.26	0.91 ± 0.06	2.65 ± 0.32	0.81 ± 0.04	1.37 ± 0.38	0.91 ± 0.07
	IW-CP	1.31 ± 0.75	0.86 ± 0.12	1.31 ± 0.73	0.60 ± 0.14	1.05 ± 0.45	0.79 ± 0.14
AaD	CP (oracle)	1.08 ± 0.18	0.91 ± 0.06	5.15 ± 0.75	0.90 ± 0.03	1.07 ± 0.22	0.91 ± 0.07
	SFCP	1.05 ± 0.16	0.91 ± 0.06	2.36 ± 0.26	0.83 ± 0.03	1.07 ± 0.22	0.91 ± 0.07
	IW-CP	1.07 ± 0.52	0.85 ± 0.11	1.38 ± 0.45	0.70 ± 0.09	0.94 ± 0.38	0.80 ± 0.14

Compared methods. We implemented the conformal prediction scores RAPS (Angelopoulos et al., 2021) and HPS (Vovk et al., 2005). We compared the following three CP methods: (1) CP (oracle) - using a validation set from the target domain with the true labels, (2) IW-CP (Tibshirani et al., 2019) - using labeled data from the source domain and unlabeled data from the target domain, (3) SFCP (our approach) - using unlabeled data from the target domain.

Evaluation Measures. We evaluated each CP method based on the average size of the prediction sets (where a small value means high efficiency) and the fraction of test samples for which the prediction sets contained the true labels. The two evaluation metrics are formally defined as:

$$\text{size} = \frac{1}{n}\sum_i |C(x_i)|, \quad \text{coverage} = \frac{1}{n}\sum_i 1(y_i \in C(x_i))$$

such that n is the size of the test set. We report the mean and standard deviation over 1000 random splits of validation and test points.

Datasets. We report experiments on the following standard domain adaptation benchmarks: Office-Home (Venkateswara et al., 2017), Office-31 (Saenko et al., 2010), VisDA (Peng et al., 2017), and DomainNet (Peng et al., 2019). Office-home is a dataset that contains 4 domains where each domain consists of 65 categories. The four domains are: Art (A) – artistic images in the form of sketches, paintings, ornamentation, etc.; Clipart (C) – a collection of clipart images; Product (P) – images of objects without a background and

Table 4: RAPS calibration results for $1-\alpha = 0.9$ on **Office-Home**, across various SFDA classification tasks and methods with different CP methods. We report the mean and the std over 1000 different splits.

SFDA	Method	A→C		A→P		A→R		C→A	
		size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)
DCPL	CP (oracle)	10.73 ± 1.05	0.90 ± 0.03	2.79 ± 0.30	0.90 ± 0.02	2.32 ± 0.26	0.90 ± 0.03	5.22 ± 0.57	0.91 ± 0.03
	SFCP	5.20 ± 0.42	0.85 ± 0.03	1.96 ± 0.21	0.86 ± 0.03	2.05 ± 0.22	0.88 ± 0.03	4.31 ± 0.53	0.87 ± 0.03
	IW-CP	2.63 ± 0.50	0.73 ± 0.05	1.37 ± 0.41	0.73 ± 0.10	1.92 ± 0.31	0.86 ± 0.04	2.48 ± 0.72	0.74 ± 0.07
SHOT	CP (oracle)	15.95 ± 0.93	0.90 ± 0.03	4.36 ± 0.38	0.90 ± 0.03	3.10 ± 0.31	0.90 ± 0.02	6.67 ± 0.64	0.91 ± 0.03
	SFCP	7.23 ± 0.48	0.83 ± 0.03	2.63 ± 0.28	0.84 ± 0.03	2.56 ± 0.28	0.87 ± 0.03	5.30 ± 0.57	0.87 ± 0.03
	IW-CP	2.70 ± 0.69	0.62 ± 0.06	1.66 ± 0.46	0.72 ± 0.07	2.28 ± 0.39	0.84 ± 0.04	2.75 ± 0.88	0.66 ± 0.08
AaD	CP (oracle)	22.20 ± 1.39	0.90 ± 0.02	4.69 ± 0.42	0.90 ± 0.02	2.96 ± 0.28	0.90 ± 0.02	8.76 ± 1.03	0.91 ± 0.03
	SFCP	11.99 ± 0.83	0.86 ± 0.03	2.42 ± 0.25	0.85 ± 0.03	2.44 ± 0.24	0.87 ± 0.03	6.04 ± 0.54	0.87 ± 0.04
	IW-CP	2.22 ± 0.47	0.59 ± 0.07	3.49 ± 0.33	0.83 ± 0.04	1.94 ± 0.34	0.82 ± 0.04	2.94 ± 0.89	0.68 ± 0.10

SFDA	Method	C→P		C→R		P→A		P→C	
		size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)
DCPL	CP (oracle)	2.54 ± 0.29	0.90 ± 0.03	2.57 ± 0.27	0.90 ± 0.03	5.75 ± 0.64	0.90 ± 0.03	14.00 ± 0.87	0.90 ± 0.03
	SFCP	2.07 ± 0.23	0.87 ± 0.02	2.30 ± 0.24	0.88 ± 0.03	4.95 ± 0.60	0.88 ± 0.03	5.65 ± 0.35	0.83 ± 0.03
	IW-CP	1.42 ± 0.35	0.77 ± 0.07	1.81 ± 0.47	0.81 ± 0.06	2.59 ± 1.01	0.73 ± 0.09	3.06 ± 0.35	0.72 ± 0.04
SHOT	CP (oracle)	3.93 ± 0.35	0.90 ± 0.02	3.91 ± 0.34	0.90 ± 0.03	7.45 ± 0.68	0.91 ± 0.03	19.99 ± 1.58	0.90 ± 0.03
	SFCP	2.95 ± 0.30	0.86 ± 0.03	3.14 ± 0.33	0.87 ± 0.03	5.98 ± 0.66	0.87 ± 0.03	7.45 ± 0.46	0.81 ± 0.03
	IW-CP	2.07 ± 0.54	0.79 ± 0.05	2.19 ± 0.57	0.79 ± 0.05	3.25 ± 1.37	0.72 ± 0.10	2.79 ± 0.69	0.61 ± 0.05
AaD	CP (oracle)	5.14 ± 0.39	0.90 ± 0.02	3.96 ± 0.34	0.90 ± 0.02	9.37 ± 0.60	0.90 ± 0.03	22.48 ± 1.15	0.90 ± 0.02
	SFCP	3.39 ± 0.30	0.87 ± 0.03	3.06 ± 0.28	0.87 ± 0.03	7.74 ± 0.57	0.88 ± 0.03	12.90 ± 1.17	0.85 ± 0.03
	IW-CP	1.77 ± 0.40	0.76 ± 0.05	2.54 ± 0.32	0.84 ± 0.03	2.17 ± 0.68	0.63 ± 0.07	3.26 ± 0.41	0.67 ± 0.04

SFDA	Method	P→R		R→A		R→C		R→P	
		size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)
DCPL	CP (oracle)	2.65 ± 0.29	0.90 ± 0.03	4.60 ± 0.56	0.91 ± 0.03	11.51 ± 1.02	0.90 ± 0.03	1.92 ± 0.22	0.90 ± 0.03
	SFCP	2.19 ± 0.24	0.87 ± 0.03	3.63 ± 0.49	0.85 ± 0.04	4.75 ± 0.39	0.83 ± 0.03	1.75 ± 0.20	0.88 ± 0.03
	IW-CP	1.80 ± 0.38	0.82 ± 0.04	2.95 ± 0.52	0.81 ± 0.05	2.28 ± 0.46	0.71 ± 0.04	1.57 ± 0.23	0.86 ± 0.03
SHOT	CP (oracle)	3.23 ± 0.34	0.90 ± 0.03	6.06 ± 0.66	0.90 ± 0.03	15.19 ± 1.29	0.90 ± 0.02	2.61 ± 0.28	0.90 ± 0.03
	SFCP	2.69 ± 0.29	0.87 ± 0.03	4.33 ± 0.57	0.85 ± 0.04	6.10 ± 0.43	0.82 ± 0.03	2.16 ± 0.24	0.87 ± 0.03
	IW-CP	2.05 ± 0.38	0.81 ± 0.04	3.24 ± 0.78	0.79 ± 0.06	3.37 ± 0.52	0.71 ± 0.04	1.72 ± 0.25	0.82 ± 0.04
AaD	CP (oracle)	3.05 ± 0.28	0.90 ± 0.03	6.38 ± 0.59	0.90 ± 0.03	18.43 ± 1.42	0.90 ± 0.03	2.55 ± 0.27	0.90 ± 0.03
	SFCP	2.63 ± 0.25	0.88 ± 0.03	4.72 ± 0.53	0.86 ± 0.04	11.67 ± 0.84	0.86 ± 0.03	1.99 ± 0.20	0.87 ± 0.03
	IW-CP	1.82 ± 0.31	0.80 ± 0.04	2.64 ± 0.52	0.75 ± 0.05	3.22 ± 0.31	0.70 ± 0.03	1.51 ± 0.19	0.81 ± 0.04

Table 5: RAPS calibration results for $1-\alpha = 0.9$ on **VisDA**, across various SFDA classification tasks and methods with different CP methods.

Method	DCPL		SHOT		AaD	
	size ↓	coverage (%)	size ↓	coverage (%)	size ↓	coverage (%)
CP (oracle)	1.61 ± 0.04	0.90 ± 0.01	2.31 ± 0.07	0.90 ± 0.01	1.28 ± 0.02	0.90 ± 0.01
SFCP	2.43 ± 0.07	0.92 ± 0.01	3.05 ± 0.07	0.92 ± 0.01	2.58 ± 0.06	0.94 ± 0.00
IW-CP	1.01 ± 0.17	0.76 ± 0.08	0.85 ± 0.21	0.61 ± 0.10	0.86 ± 0.16	0.69 ± 0.11

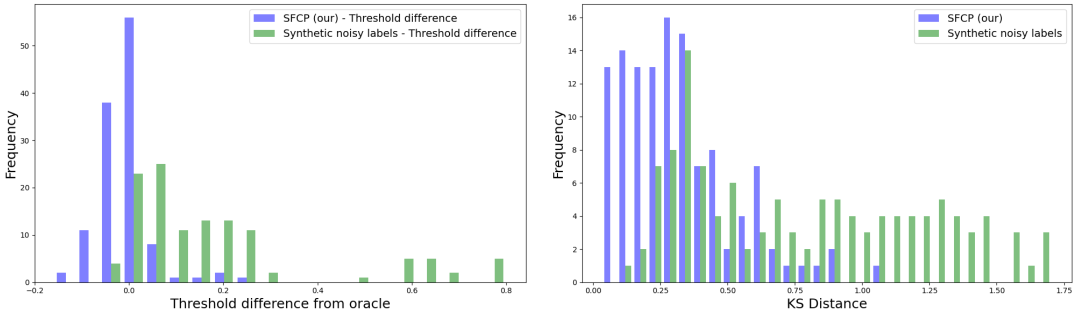

Figure 3: (left) A bar plot of CP threshold differences $\tilde{q} - q$ from the oracle across all the evaluated domain pairs for SFCP pseudo-labels (blue) and synthetic noisy labels (green). (right) A bar plot of KS statistic of the distribution functions for true labels and noisy labels across all domain pairs, for SFCP pseudo-labels (blue) and synthetic noisy labels (green).

Real-World (R) – images of objects captured with a regular camera. Office-31 is a dataset that contains 31 object categories across three domains: Amazon (A), DSLR (D), and Webcam (W). These categories include common office items such as keyboards, file cabinets, and laptops.VisDA is a simulation-to-real dataset for domain adaptation with over 280,000 images across 12 categories. DomainNet is a large UDA dataset featuring common objects. The full dataset has 345 classes, but due to labeling noise in the complete version, we used two subsets: one with 126 classes (Zhang et al., 2023; Diamant et al., 2024) and the other with 40 classes (Tan et al., 2020; Diamant et al., 2024). We refer to these subsets as DomainNet126 and DomainNet40. Both subsets included four distinct domains: Clipart (C), Product (P), Real (R) and Sketch (S) images. Overall, we utilized 102 source-target pairs. There were 12 pairs from Office-Home, 1 from VisDA, 12 from DomainNet40, and 9 from DomainNet126. Since we applied three SFDA methods, the total number of source-target pairs was (12+1+12+9) ×3 = 102.

Implementation Details. In our experiments, we employed three SFDA methods: DCPL (Diamant et al., 2024), SHOT (Jian Liang, 2020), and AaD (Yang et al., 2022) training all models to convergence using their official implementations. The three SFDA methods SHOT, AaD and DCPL represent three different approaches with respect to our generated pseudo-labels based calibration method. DCPL uses the same pseudo-labels without changing them during training, SHOT uses pseudo-labels that rely only on the source model and not on a strong pre-trained network (and also adapted during training) and AaD, does not use pseudo-labels at all in the target adaptation training. Each dataset was tested using three different random seeds, and we reported the average results. Each target domain was split into 80% for training and 20% for validation. Adaptation was performed on the training set. For the validation set, we created 1000 additional splits, with 80% of each split used for calibration and 20% used for testing. We report the mean

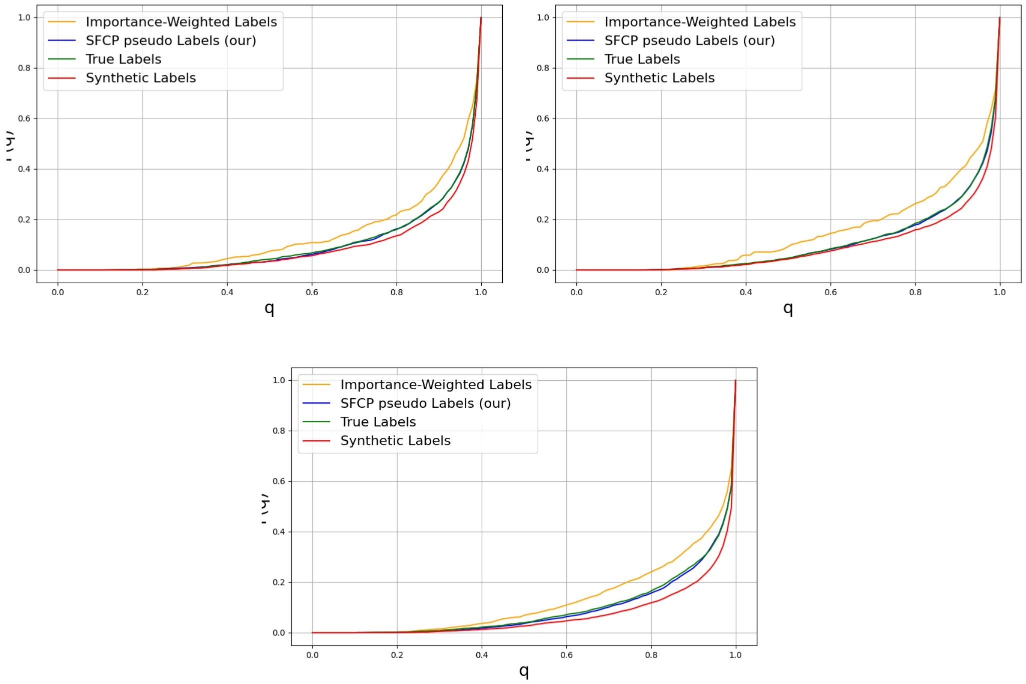

Figure 4: Plots for the APS based CDF function $F(q)$ for correct label, pseudo-labels, IW based labels and synthetically-noisy labels.

and the std over the 1000 different splits. For reproducibility, we have made our code available [1].

CP Results. Tables 1, 2, 3, 4 and 5 present the calibration results for DomainNet40, DomainNet126, Office-31, Office-Home, and VisDA respectively. The findings show that SFCP yields a coverage closer to $(1-\alpha)$ than the IW method on nearly all tasks, despite the latter having access to source domain labeled data. SFCP achieved good results for both SFDA methods based on pseudo-labels (DCPL and SHOT) and those treating the SFDA problem as an unsupervised clustering problem (AaD). A summary of the coverage across all the evaluated domain pairs is presented in Fig. 2. It shows that SFCP achieved coverage values that were concentrated around $1-\alpha$ (0.9), demonstrating its consistency. In contrast, the IW-CP method exhibited significant variability, spreading coverage values widely across a broader range. Fig. 2 also shows the same results for the HPS (Vovk et al., 2005) conformal prediction score.

Pseudo-label model analysis. We next empirically show that the threshold computed by the pseudo-labels is indeed close to the true threshold. Let q and $\tilde{q}$ respectively be the CP threshold based on the true and pseudo-labels. Fig. 3a presents a bar plot of $\tilde{q} - q$ across the more than 100 source-target domain pairs used in our experiments. We also show results for synthetic noisy labels generated using a noise transition matrix derived

[1] https://anonymous.4open.science/r/SFCP-531F

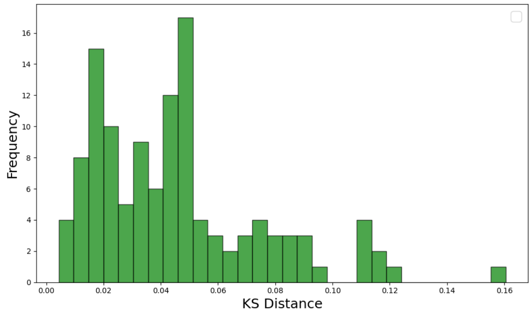

Figure 5: Bar plot of the KS distance $\Delta = \mathrm{KS}(F, \tilde{F})$ across all the evaluated domain pairs using RAPS.

from the conditional probabilities $p(\tilde{y} \mid y)$ where $\tilde{y}$ represents the pseudo-labels given the true labels y. Despite the inherent noise in the pseudo-labels, their use for CP threshold estimation results in $\tilde{q} - q$ values that are close to zero in most cases. In contrast, threshold estimations based on synthetic noisy labels consistently exhibit positive deviations from the true threshold. This makes the CP algorithm in the case of synthetic noisy labels very conservative which is aligned with observations in (Einbinder et al., 2024). We also note that since the true labels are unknown, in practice we cannot compute the distribution $p(\tilde{y}|y)$.

Fig. 4 shows plots of the function $F(q) = p(y \in C_q(x))$ for three examples of source-target domain pairs. We computed $F(q)$ for the cases of noise-free, IW method, synthetic pseudo-labels and real pseudo-labels (our approach). We can see that indeed the plots of the noise-free and pseudo-label cases are almost identical and their KS distance is small. Fig. 5 shows the statistics of $\Delta = \mathrm{KS}(F, \tilde{F})$ across all the evaluated source-target domain pairs. We see that indeed Δ is small and thus the coverage guarantee is informative.

We next empirically validate that the finite-set estimations of $F(q)$ and $\tilde{F}(q)$ pass the Kolmogorov–Smirnov (KS) test, i.e. there is no statistically significant evidence to reject the hypothesis that the data samples for $S(x, y)$ and from $S(x, \tilde{y})$ comes from the same distribution. Given a labeled validation set, we define the empirical CDFs: $G(q) = \frac{1}{n}\sum_i 1_{\{S(x_i,y_i) \le q\}}$ and $\tilde{G}(q) = \frac{1}{n}\sum_i 1_{\{S(x_i,\tilde{y}_i) \le q\}}$. Fig. 3b shows the KS distance $T = \sqrt{\frac{n}{2}}\mathrm{KS}(G, \tilde{G})$ for all the evaluated domain pairs. The null hypothesis here is that the labels y and the pseudo-labels $\tilde{y}$ are sampled from the same distribution. The KS test rejects the null hypothesis at level 90% if $T > 1.22$. We can see that the hypothesis that pseudo-labels and true labels have the same distribution passed the KS test in all 102 source-target domain pairs we checked. This was not the case for simulated, synthetic noisy labels.

5. Conclusions

In this work, we tackled the challenge of applying conformal prediction (CP) to models adapted to a target domain without access to labeled target data or source domain samples. Our extensive empirical evaluation demonstrated that the commonly used Importance Weighting (IW) method performs poorly in practice—even when source labels are available—and is not a viable solution for SFDA. In contrast, we proposed a simple yet effective approach, Source-Free Conformal Prediction (SFCP), that calibrates the adapted model using pseudo-labels generated by the source model. Despite their inherent noise, we have shown both theoretically and empirically that pseudo-labels can be surprisingly effective for CP calibration, achieving coverage levels close to those obtained using true labels. Our method consistently outperformed IW-based approaches across more than 100 domain shifts, without requiring hyperparameter tuning or access to source data. These results indicate that pseudo-label-based calibration offers a practical and reliable strategy for uncertainty quantification in challenging SFDA settings, enabling broader deployment of CP techniques under real-world domain shifts. Future directions include extending SFCP to regression tasks, sequential adaptation, and the development of improved pseudo-label refinement strategies to enhance calibration under severe shifts. While our method performs well under moderate domain shifts, we also observe limitations when the domain gap is too large, as pseudo-labels become unreliable. Addressing this limitation is a promising direction for future research.

References

Jiahao Ai and Zhimei Ren. Not all distributional shifts are equal: Fine-grained robust conformal inference. *arXiv preprint arXiv:2402.13042*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *International Conference on Learning Representations (ICLR)*, 2021.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C Duchi. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, 119(548):3033–3044, 2024.

Weijie Chen, Luojun Lin, Shicai Yang, Di Xie, Shiliang Pu, and Yueting Zhuang. Self-supervised noisy label learning for source-free unsupervised domain adaptation. In *International Conference on Intelligent Robots and Systems (IROS)*, 2022.

Jase Clarkson, Wenkai Xu, Mihai i Cucuringu, and Gesine Reinert. Split conformal prediction under data contamination. In *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, 2024.

Idit Diamant, Amir Rosenfeld, Idan Achituve, Jacob Goldberger, and Arnon Netzer. De-confusing pseudo-labels in source-free domain adaptation. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2024.

Bat-Sheva Einbinder, Shai Feldman, Stephen Bates, Anastasios N Angelopoulos, Asaf Gendler, and Yaniv Romano. Label noise robustness of conformal prediction. *Journal of Machine Learning Research*, 25(328):1–66, 2024.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Jiawei Ge, Debarghya Mukherjee, and Jianqing Fan. Optimal aggregation of prediction intervals under unsupervised domain shift. *arXiv preprint arXiv:2405.10302*, 2024.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2021.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems (NeurIPs)*, 2021.

Isaac Gibbs and Emmanuel J Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Pierre Humbert, Batiste Le Bars, Aurélien Bellet, and Sylvain Arlot. One-shot federated conformal prediction. In *International Conference on Machine Learning (ICML)*, 2023.

Jiashi Feng Jian Liang, Dapeng Hu. Do we really need to access the source data? In *International Conference on Machine Learning (ICML)*, 2020.

N. Karim, N. Mithun, A. Rajvanshi, H. Chiu, S. Samarasekera, and N. Rahnavard. C-SFDA: A curriculum learning aided self-training framework for efficient source free domain adaptation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *International Conference on Machine Learning (ICML)*, 2021.

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 (ICML)*, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

Charles Lu, Anastasios N Angelopoulos, and Stuart Pomerantz. Improving trustworthiness of AI disease severity rating in medical imaging with ordinal conformal prediction sets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2022a.

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022b.

Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In *International Conference on Machine Learning (ICML)*, 2023.

Henrik Olsson, Kimmo Kartasalo, Nita Mulliqi, et al. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. *Nature Communications*, 13(1):7761, 2022.

Anusri Pampari and Stefano Ermon. Unsupervised calibration under covariate shift. *arXiv preprint arXiv:2006.16405*, 2020.

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 *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. Conformal prediction for federated uncertainty quantification under label shift. In *International Conference on Machine Learning (ICML)*, 2023.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems (NeurIPs)*, 2020.

Aaron Roth. Uncertain: Modern topics in uncertainty estimation. *Unpublished Lecture Notes*, 2022. URL https://www.cis.upenn.edu/~aaroth/uncertainty-notes.pdf.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2010.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Matteo Sesia, YX Rachel Wang, and Xin Tong. Adaptive conformal classification with noisy labels. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2024.

Shuhan Tan, Xingchao Peng, and Kate Saenko. Class-imbalanced domain adaptation: An empirical odyssey. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2020.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPs)*, 2019.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable calibration with lower bias and variance in domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPs)*, 2020.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning (ICML)*, 2021.

Ge Yan, Yaniv Romano, and Tsui Weng. Provably robust conformal prediction with improved efficiency. In *International Conference on Learning Representations (ICLR)*, 2024.

Shiqi Yang, Yaxing Wang, Kai Wang, Shangling Jui, and Joost van de Weijer. Attracting and dispersing: A simple approach for source-free domain adaptation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Li Yi, Gezheng Xu, Pengcheng Xu, Jiaqi Li, Ruizhi Pu, Charles Ling, A. Ian McLeod, and Boyu Wang. When source-free domain adaptation meets learning with noisy labels. In *International Conference on Learning Representations (ICLR)*, 2023.

Wenyu Zhang, Li Shen, and Chuan-Sheng Foo. Rethinking the role of pre-trained networks in source-free domain adaptation. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

Xin Zou and Weiwei Liu. Coverage-guaranteed prediction sets for out-of-distribution data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2019.