REVISITING SOURCE-FREE DOMAIN ADAPTATION: A NEW PERSPECTIVE VIA UNCERTAINTY CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

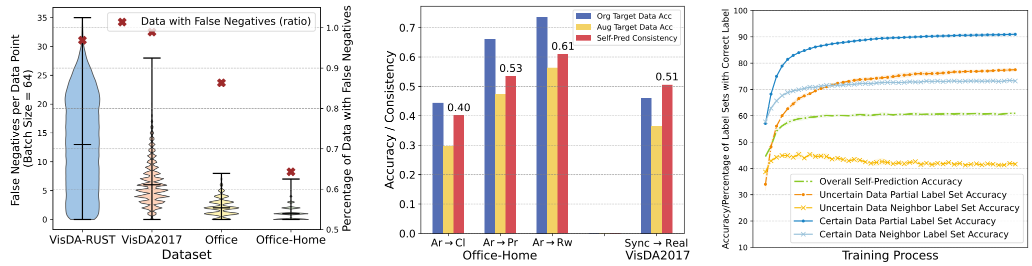
Source-Free Domain Adaptation (SFDA) seeks to adapt a pre-trained source model to the target domain using only unlabeled target data, without access to the original source data. While current state-of-the-art (SOTA) methods rely on leveraging weak supervision from the source model to extract reliable information for self-supervised adaptation, they often overlook the uncertainty that arises during the transfer process. In this paper, we conduct a systematic and theoretical analysis of the uncertainty inherent in existing SFDA methods and demonstrate its impact on transfer performance through the lens of Distributionally Robust Optimization (DRO). Building upon the theoretical results, we propose a novel instance-dependent uncertainty control algorithm for SFDA. Our method is designed to quantify and exploit the uncertainty during the adaptation process, significantly improving the model performance. Extensive experiments on benchmark datasets and empirical analyses confirm the validity of our theoretical findings and the effectiveness of the proposed method. This work offers new insights into understanding and advancing SFDA performance.

1 INTRODUCTION

Deep neural networks (DNNs) have achieved remarkable performance across a wide range of tasks. However, their performance can experience significant declines when there is a domain shift between training (source) and test (target) data. Traditional solutions leverage transferable knowledge from labeled source data to classify unlabeled target data. However, access to source data is often restricted due to privacy concerns or proprietary constraints. To address these challenges, Source-Free Domain Adaptation (SFDA) has emerged, aiming to adapt a pre-trained source model to an unlabeled target domain without accessing the original source data (Liang et al., 2020; Yang et al., 2021b;a).

Recent work has explored the integration of self-supervised learning with transfer learning in SFDA, where contrastive learning (CL)-based self-supervised methods have gained widespread use and empirical support (Yang et al., 2022; Karim et al., 2023; Chen et al., 2022; Hwang et al., 2024; Mitsuzumi et al., 2024a). A key challenge in applying CL methods to SFDA lies in selecting and utilizing positive and negative samples of target data with a well-trained source model. Different from conventional CL methods using data augmentations as positive samples, in SFDA, the neighbors in the feature space can provide stronger supervision and usually be treated as positives, and the negative samples are the remaining data in the training mini-batch. However, due to the domain shift, these methods face severe uncertainty, as will be elaborated shortly.

In this paper, we systematically and theoretically examine the uncertainty present in SFDA through the lens of Distributionally Robust Optimization (DRO). Unlike previous studies that primarily focus on empirical strategies (Roy et al., 2022; Litrico et al., 2023; Pei et al., 2023; Lee et al., 2022), our work offers a comprehensive analysis of two types of uncertainty arising from the use of negative and positive samples in existing SFDA methods, aiming to enhance the SFDA performance through uncertainty control. Specifically, on one hand, random sampling of negative samples in practice often introduces outliers, or ‘false negative examples’ – samples that belong to the same class as the considered target data point but are mistakenly selected as negatives (as shown in Figure 1a). This leads to a deviation of the empirical negative distribution from the true distribution, thus introducing uncertainty into the loss calculation. To address this sampling bias, we define a *negative uncertainty set*, which consists of distributions obtained by slightly perturbing the training negative distribution,



(a) Negative Sampling Uncertainty (b) Inconsistency Prediction (c) Positive Supervision Uncertainty

Figure 1: *(a) Clear presence of false negative samples across different datasets. (b) Inconsistency between the prediction results for the anchor image and its augmented view by the source model. (c) Illustration of varying predictive accuracies between certain and uncertain target data during the adaptation process on Office-Home (Ar $\rightarrow$ Cl).*

and consider an outlier-robust worst-case risk within this set. We theoretically derive an upper bound for this risk, which motivates incorporating a dispersion control term into the loss function. Moreover, inspired by the prediction inconsistency phenomenon between a target image and its augmented view (Figure 1b), we propose an augmentation-based dispersion control approach to mitigate the uncertainty introduced by noisy negative samples. On the other hand, domain shift causes models trained on source data to produce uncertain probabilities when applied to target data. In such cases, the supervisory information from positive examples may not fully align with the ground truth, making the use of neighboring predictions for supervision introduce additional uncertainty. Unlike existing methods that focus on mitigating uncertainty (Roy et al., 2022; Litrico et al., 2023; Mitsuzumi et al., 2024a), we aim to utilize this information more effectively. To better accommodate the uncertainty in the predicted probabilities of positive samples, we consider a *positive uncertainty set* centered around these probabilities and examine the worst-case risk within this set. We theoretically show that the optimal solution for the target points consists of a partial label set. To make the most of this uncertain information, we propose novel criteria to identify uncertain data and use partial labels to relax supervision of these samples. As shown in Figure 1c, leveraging such uncertainty information leads to greater performance gains compared to using only certain data.

Our contributions are as follows: (1) We theoretically analyze two sources of uncertainty in contrastive learning-based SFDA methods, leading to the identification of two types of worst-case risks under a unified DRO framework. Through this investigation, we explain why current contrastive learning methods can significantly boost SFDA performance (Section 4.2) while revealing the overlooked uncertain information in existing algorithms (Section 4.3). Our theoretical analysis also provides a novel perspective in understanding the SFDA problem. (2) Based on our theoretical result, we design a novel uncertainty control algorithm for SFDA (UCon-SFDA), which minimizes the negative effect introduced by the uncertainty from negative sample distribution while leveraging the uncertain information in positive example predictions to enhance the model's discriminability (Section 4.4). (3) We conduct extensive experiments to validate the effectiveness of the proposed method.

2 Related Work

Source-Free Domain Adaptation (SFDA). SFDA focuses on adapting a well-trained source model to a target domain where only unlabeled data are available. Since source data are not accessible during adaptation, some methods rely on extracting source information through prototype generation (Qiu et al., 2021), or minimizing dependence on the source through adversarial training (Li et al., 2020b). Addressing the lack of target labels, several methods aim to obtain more accurate supervision for the target data. For example, SHOT (Liang et al., 2020) employs deep clustering to create pseudo-labels, while NRC (Yang et al., 2021a) and G-SFDA (Yang et al., 2021b) leverage neighboring predictions to guide the adaptation process. Recently, self-supervised learning has been increasingly integrated with transfer learning in SFDA, and contrastive learning-based self-supervised methods have been widely utilized and empirically validated. For instance, AaD (Yang et al., 2022) introduces positive and negative samples into SFDA and uses a simplified contrastive loss to enhance model discriminability while maintaining diversity; C-SFDA (Karim et al., 2023) utilizes a teacher-student framework to enhance the self-training in SFDA; methods like DaC (Zhang et al., 2022), AdaContrast (Chen et al., 2022), and SF(DA)^2 (Hwang et al., 2024) explore explicit or implicit data augmentation to further boost SFDA performance. I-SFDA (Mitsuzumi et al., 2024a) offers a new perspective by approaching SFDA through self-training. Despite these advancements, a comprehensive theoretical framework explaining their effectiveness is still missing. Moreover, most existing methods do not fully account for the uncertainty inherent in the adaptation process.

Uncertainty in SFDA. Given the absence of both source data and target labels, handling uncertainty is a key challenge in SFDA, especially when faced with domain shifts. Current research mostly addresses prediction or representation uncertainty by reweighting loss functions or prioritizing more confident samples during training (Roy et al., 2022; Litrico et al., 2023; Pei et al., 2023; Lee et al., 2022). In contrast to these approaches, our approach provides a systematic and comprehensive analysis of various sources of uncertainty in contrastive learning-based SFDA from the instance-dependant perspective. Building on this analysis, we propose a novel algorithm that improves SFDA performance by effectively controlling variance during adaptation.

3 Preliminaries

Notations. We use $[k]$ to denote the set $\{1, \dots, k\}$ for any positive integer k. For $a \in \mathbb{R}$, we define $a_+ = \max\{a, 0\}$, and let $\lfloor a \rfloor$ and $\lceil a \rceil$ denote the floor and the ceiling of a, respectively. For a vector $\boldsymbol{v}$, the jth element is represented as v_j, and $\boldsymbol{v}^\top$ indicates its transpose. Given $\boldsymbol{v} = (v_1, \dots, v_p)^\top$ and $q \in [1, +\infty]$, the L^q norm is defined as $\|\boldsymbol{v}\|_q = \left(\sum_{j=1}^p |v_j|^q\right)^{1/q}$ for $1 \le q < \infty$, and $\|\boldsymbol{v}\|_\infty = \max_j |v_j|$ when $q = +\infty$. Let $(\Omega, \mathcal{G}, \mu)$ represent a measure space, where Ω is a set, $\mathcal{G}$ is the σ-algebra of subsets of Ω, and μ is the associated measure. For $q > 0$, let $L^q(\Omega, \mathcal{G}, \mu)$, or simply $L^q(\mu)$, denote the space of Borel-measurable functions $f : \Omega \to \mathbb{R}$ such that $\int |f|^q d\mu < \infty$. We denote the expectation and variance of $f(\mathrm{Z})$ with respect to $\mathrm{Z} \sim \mu$ as $\mathbb{E}_\mu\{f(\mathrm{Z})\}$ and $\mathbb{V}_\mu\{f(\mathrm{Z})\}$, respectively; and when the context is clear, we simplify the notations to $\mathbb{E}_\mu(f)$ and $\mathbb{V}_\mu(f)$, respectively. We use $\mathcal{P}(\Omega)$ to denote the set of Borel probability measures on Ω, and let $\mathcal{P}_q(\Omega)$ represent the subset of $\mathcal{P}(\Omega)$ with finite qth moment for $q > 0$. That is, $\mu \in \mathcal{P}_q(\Omega)$ if and only if $\mathbb{E}_{\mathrm{Z}\sim\mu}(\mathrm{Z}^q) < \infty$.

Problem Setup. For a K-class classification problem, let $\mathcal{X} \subset \mathbb{R}^d$ represent the input space, and let $\mathcal{Y} = [K]$ denote the label space, with d denoting the input dimension. In Source-Free Domain Adaptation (SFDA), we assume that the source domain distribution $P^{\mathrm{S}}_{\mathbf{x}\mathrm{y}}$ and the target domain distribution $P^{\mathrm{T}}_{\mathbf{x}\mathrm{y}}$ are two distinct, unknown distributions over $\mathcal{X} \times \mathcal{Y}$. We express these distributions as $P^{\mathrm{S}}_{\mathbf{x}\mathrm{y}} = P^{\mathrm{S}}_{\mathbf{x}} P^{\mathrm{T}}_{\mathrm{y}|\mathbf{x}}$ and $P^{\mathrm{T}}_{\mathbf{x}\mathrm{y}} = P^{\mathrm{T}}_{\mathbf{x}} P^{\mathrm{T}}_{\mathrm{y}|\mathbf{x}}$, where the subscripts indicate the involved variables. For the source domain, we have a *source model* $h_{\mathrm{S}} : \mathcal{X} \to \mathcal{Y}$, which is a neural network-based predictor pre-trained with N_{S} labeled examples $\mathcal{D}_{\mathrm{S}} \triangleq \{\mathbf{x}^{\mathrm{S}}_i, \mathrm{y}^{\mathrm{S}}_i\}_{i=1}^{N_{\mathrm{S}}}$ drawn from $P^{\mathrm{S}}_{\mathbf{x}\mathrm{y}}$. In the target domain, let $\mathcal{D}_{\mathrm{T}} \triangleq \{\mathbf{x}^{\mathrm{T}}_i\}_{i=1}^{N_{\mathrm{T}}}$ denote the unlabeled target domain data of size N_{T}, consisting of observations of independent and identically distributed (i.i.d.) random variables drawn from $P^{\mathrm{T}}_{\mathbf{x}}$. Given the source model h_{S} and unlabeled target data $\mathcal{D}_{\mathrm{T}}$, our goal is to learn a *target model* $h_{\mathrm{T}} : \mathcal{X} \to \mathcal{Y}$ that predicts the labels in the target domain by adapting h_{S} on $\mathcal{D}_{\mathrm{T}}$.

To facilitate our analysis in the context of deep learning, we define the target model h_{T} as $h_{\mathrm{T}}(\mathbf{x}; \boldsymbol{\theta}_{\mathrm{T}}) = \arg\max_{j\in[K]} f_{\mathrm{T}}(\mathbf{x}; \boldsymbol{\theta}_{\mathrm{T}})[j]$ for any $\mathbf{x} \in \mathcal{X}$. Here, $\boldsymbol{\theta}_{\mathrm{T}} \in \boldsymbol{\Theta}$ represents the vector of model parameters in the parameter space $\boldsymbol{\Theta}$. The function $f_{\mathrm{T}} : \mathcal{X} \to \Delta^{K-1}$ denotes the network output, where Δ^{K-1} is the K-simplex, and $f_{\mathrm{T}}(\mathbf{x}; \boldsymbol{\theta}_{\mathrm{T}})[j]$ refers to the jth component of the vector-valued function f_{T}. The source model h_{S} is defined similarly, with the corresponding network $f_{\mathrm{S}}(\cdot; \boldsymbol{\theta}_{\mathrm{S}})$.

4 Theoretical Analysis and Algorithm

4.1 Motivation

Existing SFDA methods typically decompose their training loss into two components: discriminability, which enhances the model's ability to distinguish between unlabeled target samples, and diversity, which promotes predictions across diverse classes (Yang et al., 2022; Mitsuzumi et al., 2024b; Cui et al., 2020). Among these, contrastive learning methods are the most widely used, where the goal is to maximize the similarity between positive pairs to improve discriminability and minimize the similarity between negative pairs to ensure diversity. This can be formulated as the following expected risk with contrastive loss:

$$\mathcal{R}_{\text{basic}}(\boldsymbol{\theta}) = \mathbb{E}_{P^{\mathrm{T}}_{\mathbf{x}}}\left[-\mathbb{E}_{P^{+}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{+};\mathbf{X})\right\} + \mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{X})\right\}\right], \tag{1}$$

where the outer expectation $\mathbb{E}_{P^{\mathrm{T}}_{\mathbf{x}}}$ is taken over the input data distribution $\mathbf{X}$, while the inner expectations $\mathbb{E}_{P^{+}}$ and $\mathbb{E}_{P^{-}}$ are evaluated under the conditional distributions of positive example $\mathbf{X}^{+}$ and negative example $\mathbf{X}^{-}$, respectively, given $\mathbf{X}$. Here, function $\mathcal{S}_{\boldsymbol{\theta}}(\cdot;\cdot)$, mapping from $\mathcal{X}\times\mathcal{X}$ to $[0,1]$, represents the similarity measure between two instances, which, for instance, can be taken as the cosine similarity computed as the dot product of their corresponding network outputs.

In contrastive learning-based SFDA, for each target input $\mathbf{x}_i^{\mathrm{T}}$ in a mini-batch $\mathcal{B}$, the set of positive examples of $\mathbf{x}_i^{\mathrm{T}}$, denoted $\mathcal{C}_i$, consists of the κ-nearest neighbours in the training set $\mathcal{D}_{\mathrm{T}}$ for some positive integer κ typically chosen between 2 and 5; while the negative set is taken as $\mathcal{B}\backslash\{\mathbf{x}_i^{\mathrm{T}}\}$. However, this construction inevitably includes a fraction of false negatives, leading to sampling bias and deviation from the true underlying distribution. While with the help of a well-trained source model, neighboring positive samples in the feature space often provide effective supervision for most unlabeled target domain data, some highly uncertain samples persist due to domain shift. To address these issues, we propose a robust strategy for managing uncertainty in SFDA using distributionally robust optimization (DRO).

4.2 Negative Sampling Uncertainty and Dispersion Control

To address the uncertainty from sampling bias and distribution shift in negative examples, we consider an expected distributionally robust optimization (DRO) risk: for each given $\mathbf{x}\in\mathcal{X}$ and $\delta>0$,

$$\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta};P^{-},\delta) = \sup_{Q^{-}\in\Gamma_\delta(P^{-})}\left[\mathbb{E}_{Q^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}\right], \tag{2}$$

where the expectation $\mathbb{E}_{Q^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}$ is evaluated under the conditional distribution Q^{-} of $\mathbf{X}^{-}$, given $\mathbf{X}=\mathbf{x}$. The set $\Gamma_\delta(P^{-})$ represents an *uncertainty set* of probability measures centered around the *reference probability distribution* P^{-}, with a radius $\delta>0$ that controls the robustness (Gao, 2023; Gao et al., 2024; Duchi & Namkoong, 2021). A common way is to define $\Gamma_\delta(P^{-})$ as the distance-based uncertainty set:

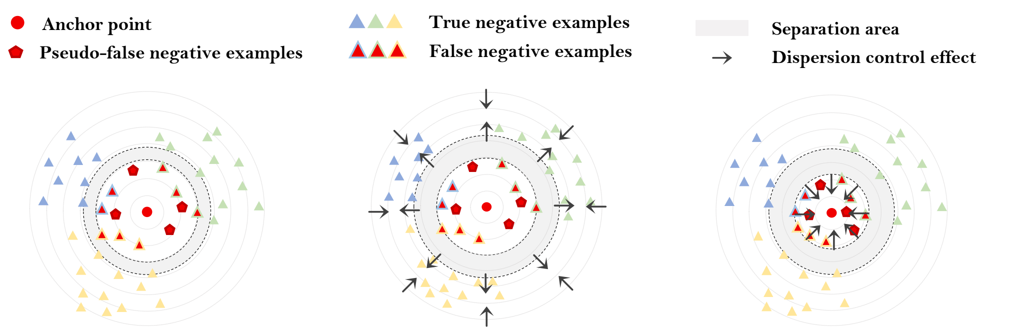
$$\Gamma_\delta(P^{-}) = \{Q^{-}\in\mathcal{P}_p(\mathcal{X}) : d\!\!\!{-}(Q^{-},P^{-})\le\delta\}, \tag{3}$$

where $\mathcal{P}_p(\mathcal{X})$ denotes the class of Borel probability measures on $\mathcal{X}$ with finite pth moment for some $p>1$, and $d\!\!\!{-}$ is a discrepancy metric of probability measures. Popular choices of $d\!\!\!{-}$ are φ-divergences (including Kullback–Leibler (KL) divergence and χ^2 divergence as special cases (Duchi, 2016)) and Wasserstein distances (Gao, 2023; Gao et al., 2024; Blanchet & Murthy, 2019).

In practice, negative samples are often drawn uniformly from the training data, often leading to the inclusion of false negatives. Let P^{-}_{train} represent the observed distribution of these negative samples, modeled using Huber's ϵ-contamination framework: $P^{-}_{\text{train}} = (1-\epsilon)P^{-} + \epsilon\widetilde{P}^{-}$, where $\epsilon\in(0,1)$ is the contamination level, and $\widetilde{P}^{-}$ represents an arbitrary contamination distribution (Huber, 1992). For instance, consider some $\mathbf{x}\in\mathcal{X}$. Suppose we collect n negative samples, where a fraction $\lfloor\varepsilon n\rfloor$ are i.i.d. false negative examples drawn from $\widetilde{P}^{-}$, and the rest are true negatives from P^{-}. The resulting empirical distribution of the observed negative samples follows this model with contamination level $\lfloor\varepsilon n\rfloor/n$. To mitigate overfitting to the worst-case instances that are likely to be outliers, we minimize the refined outlier-robust expected risk (Nietert et al., 2024a;b; Zhai et al., 2021):

$$\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta};P^{-}_{\text{train}},\delta,\epsilon) = \inf_{P'\in\mathcal{P}_p(\mathcal{X})}\left\{\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta};P',\delta) : \exists\widetilde{P}'\in\mathcal{P}_p(\mathcal{X})\ s.t.\ P^{-}_{\text{train}} = (1-\epsilon)P' + \epsilon\widetilde{P}'\right\}. \tag{4}$$

By definition, the minimizer of (4) is designed to ignore 'hard' data points that contribute most to worst-case risk, and instead focus on the $(1-\epsilon)$-fraction of 'easy' data points in the training set. This helps prevent overfitting to outliers, thereby reducing the risk of pushing the target data point away from others within the same class. For different choices of the discrepancy metric $d\!\!\!{-}$ in the uncertainty set (3), we establish a unified upper bound on the outlier-robust risk $\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta};P^{-}_{\text{train}},\delta,\epsilon)$. The result is summarized in the following informal theorem, with the formal statement and its proof provided in Appendix A.3.



(a) No Dispersion Control (b) Direct Dispersion Control (c) Dispersion Control with Pseudo-False Negatives

Figure 2: *Visualization of the effect of dispersion control. (a) No dispersion control. (b) Direct dispersion control between the anchor and false-negative pairs. (c) Dispersion control with pseudo-false negatives.*

Theorem 4.1 (informal). *Suppose the similarity measure $\mathcal{S}_{\boldsymbol{\theta}}$ satisfies the smoothness conditions in Lemma 5 for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. For the contaminated training distribution P^-_{train}, let p^-_{train} denote the associated density/mass function, and we defined the associated truncated distribution P^* with density/mass function p^*: $p^*(\mathbf{x}^-) \triangleq \frac{1}{1-\epsilon} p^-_{train}(\mathbf{x}^-)\mathbf{1}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^-;\mathbf{x}) \le s^*\right\}$, where s^* is the $1-\epsilon$ quantile satisfying $P^-_{train}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x}) \le s^*\right\} = 1-\epsilon$. Then, for a small enough $\delta > 0$, we have*

$$\mathcal{R}^-_{\mathbf{x}}(\boldsymbol{\theta}; P^-_{train}, \delta, \epsilon) \le \mathbb{E}_{P^*}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})\right\} + \mathcal{V}_d\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x}); P^*\right\} + \mathcal{O}(\delta),$$

where $\mathcal{V}_d(\cdot; P^*)$ is a measure of statistical dispersion that depends on the choice of the discrepancy metric d, and is evaluated under the truncated distribution P^*.

Remark 4.1. In contrastive SFDA, for each *anchor point* $\mathbf{x}$ from the target set (i.e., the data point we use as a reference to compare with positive and negative examples), the truncated version of P^-_{train}, denoted as P^* in Theorem 4.1, concentrates all its mass on regions where the similarity falls below the $(1-\epsilon)$-quantile. Consequently, the first expectation term in the upper bound controls the average risk over potential true negative examples, akin to the behavior of traditional negative sample loss (Yang et al., 2022; Mitsuzumi et al., 2024b). This is implemented as the negative sample loss $\mathcal{L}^-_{\text{CL}}$ in (7) presented in Section 4.4. Meanwhile, the second term of Theorem 4.1 manages the dispersion in similarity between these true negative examples, helping to distinguish the anchor-true-negative pairs from the anchor-false-negative ones. This term encourages greater separation between the prediction similarities for anchor-true-negative pairs and anchor-false-negative pairs, as shown by the wider gray area in Figure 2b than that in Figure 2a.

Remark 4.2. In practice, domain shift makes it challenging to distinguish between false negatives and true negatives. To address this, we propose to achieve dispersion control by manually constructing pseudo-false negative examples using techniques such as data augmentation. As shown in Figure 1b, for a given anchor point $\mathbf{x}$, the source model's prediction on its augmented version, denoted as $\texttt{AUG}(\mathbf{x})$, may not align with the prediction for $\mathbf{x}$. When this happens, $\texttt{AUG}(\mathbf{x})$ is automatically treated as a false negative example for $\mathbf{x}$. Motivated by the dispersion control term, we treat these augmentations as pseudo-false negatives and minimize the negative similarity between the anchor point and its augmented prediction. As illustrated in Figure 2c, this can effectively push the similarity of anchor-false-negative pairs farther from that of anchor-true-negative pairs, increasing the width of the gray region area to achieve the desired separation and dispersion control. This dispersion control effect is captured through the loss term $\mathcal{L}^-_{\text{DC}}$ in (7), as detailed in Section 4.4.

4.3 Positive Supervision Uncertainty and Partial Labeling

For each anchor point $\mathbf{x} \in \mathcal{X}$ in the target dataset, let $p \triangleq (p_1, \ldots, p_K)^\top \triangleq f_{\text{T}}(\mathbf{x}; \boldsymbol{\theta}) \in \Delta^{K-1}$ denote the target model's predicted probabilities for $\mathbf{x}$. For the positive example $\mathbf{x}^+$ associated with $\mathbf{x}$, let $p^+ \triangleq (p^+_1, \ldots, p^+_K)^\top$ represent the predicted probabilities for $\mathbf{x}^+$, which could come from a source model or previous training iterations. When using cosine similarity, the positive supervision from $\mathbf{x}^+$ encourages the model training to minimize the negative similarity, defined as $-\langle p^+, p\rangle = -\sum_{j=1}^{K} p_j p_j^+$.

In SFDA, leveraging a well-trained source model and the similarity between the source and target domain distributions, the neighboring examples in the feature space are often treated as positive samples. While many of these positive samples provide effective supervision for unlabeled target data, there can still be highly uncertain examples due to domain shift. To better handle this uncertainty in model predictions, we explore the optimal prediction for the anchor point $\mathbf{x}$ by solving the following worst-case risk minimization problem based on DRO:

$$p^\star \in \inf_{p\in\Delta^{K-1}} \mathcal{R}^+_{\mathbf{x}}(p;\mathbf{x}^+,\delta), \text{ with } \mathcal{R}^+_{\mathbf{x}}(p;\mathbf{x}^+,\delta) \triangleq \sup_{q^+\in\Gamma_\delta(p^+)} \langle q^+, -p\rangle, \tag{5}$$

where $\Gamma_\delta(p^+)$ is the uncertainty set centered around the reference distribution p^+, as defined in (3). If we use the p-Wasserstein distance (Definition A.1), with the 0-1 cost function, as the discrepancy metric in the uncertainty set, we can derive a closed-form expression for $p^\star$ as follows.

Theorem 4.2. *Let $\{p_1^+,\dots,p_K^+\}$ be arranged in decreasing order, denoted $p^+_{(1)} \geq \dots p^+_{(K)}$, with the corresponding indexes denoted $\chi(1),\dots,\chi(K)$. Let $p_{(j)}$ denote the $\chi(j)$-th component of p, corresponding to $p^+_{(j)}$ for $j\in[K]$. Then, the optimal solution $p^\star$ of (5) is given as follows:*

- *If $\frac{1}{K} \geq \frac{1}{k^*}\sum_{j=1}^{k^*} p^+_{(j)} - \frac{1}{k^*}\delta^p$ for all $k^*\in[K-1]$, then $p^\star_j = \frac{1}{K}$ for all $j\in[K]$.*
- *If there exists some $k_0\in[K-1]$ such that $\frac{1}{k_0}\sum_{j=1}^{k_0} p^+_{(j)} - \frac{1}{k_0}\delta^p > \frac{1}{K}$ and $\frac{1}{k_0}\sum_{j=1}^{k_0} p^+_{(j)} - \frac{1}{k_0}\delta^p \geq \frac{1}{k^*}\sum_{j=1}^{k^*} p^+_{(j)} - \frac{1}{k^*}\delta^p$ for all $k^*\in[K-1]$, then $p^\star_{(j)} = \frac{1}{k_0}$ for $j\in[k_0]$ and $p^\star_{(j)} = 0$ for $j = k_0+1,\dots,K$.*

Remark 4.3. Theorem 4.2 suggests that the optimal prediction for an anchor point can be represented by a set of *(instance-dependent) partial labels*. The advantage of using partial labels, rather than the entire predicted probabilities, as the supervision signal is that it retains uncertain yet potentially more accurate label information, while eliminating interference from labels that are more likely to be incorrect. In the special case where $p^+_{(1)} \geq \max\{\frac{1}{K}+\delta^p, p^+_{(2)}+\delta^p\}$, the optimal solution simplifies to $p^\star_{(1)} = 1$ and $p^\star_{(j)} = 0$ for $j = 2,\dots,K$. That is, the optimal solution is to select the label with the highest predicted probability for the anchor point, rather than a set of partial labels, when the gap between the top two probabilities exceeds a given threshold. We term this scenario *certain label information*; otherwise, we classify it as *uncertain label information*.

Remark 4.4. Motivated by Theorem 4.2 and Remark 4.3, we propose to leverage both certain and uncertain label information in distinct ways to effectively capture and utilize prediction uncertainty. Specifically, when an instance $\mathbf{x}$ receives *certain label information*, the optimal prediction for $\mathbf{x}$ corresponds to the label with the highest predicted probability. This certain supervision signal is incorporated through the *positive supervision loss* term $\mathcal{L}^+_{\mathrm{CL}}$ in (8). When *uncertain label information* is provided, the optimal prediction for $\mathbf{x}$ is expressed as a set of partial labels. Instead of relying solely on the estimated pseudo labels, we construct a *partial label set* for $\mathbf{x}$. This approach offers a more robust supervisory signal by accounting for multiple potential labels and reducing reliance on noisy single-label predictions. This information is captured through the *partial label loss* term $\mathcal{L}^+_{\mathrm{PL}}$ in (8). To distinguish between certain and uncertain label information in applications, we use the ratio of the two highest predicted probabilities, as detailed in Section 4.4.

4.4 Implementation

In our algorithm, we build upon the conventional contrastive loss commonly adopted in previous works (Yang et al., 2022; Mitsuzumi et al., 2024a):

$$\mathcal{L}_{\mathrm{CL}} \triangleq \mathcal{L}^+_{\mathrm{CL}} + \lambda^-_{\mathrm{CL}}\mathcal{L}^-_{\mathrm{CL}} \triangleq \frac{1}{N_{\mathrm{T}}}\sum_{i=1}^{N_{\mathrm{T}}}\Big\{-\sum_{\mathbf{x}_i^+\in\mathcal{C}_i}\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}_i^+;\mathbf{x}_i) + \lambda^-_{\mathrm{CL}}\sum_{\mathbf{x}_i^-\in\mathcal{B}\setminus\{\mathbf{x}_i\}}\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}_i^-;\mathbf{x}_i)\Big\}, \tag{6}$$

where similarity is computed as $\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}_i^{+/-};\mathbf{x}_i) = \langle f_{\mathrm{T}}(\mathbf{x}_i^{+/-};\boldsymbol{\theta}), f_{\mathrm{T}}(\mathbf{x}_i;\boldsymbol{\theta})\rangle$. Positive samples are the κ-nearest neighbours in the feature space from the training set $\mathcal{D}_{\mathrm{T}}$, and negative samples are the remaining data in the same mini-batch $\mathcal{B}$. Building on this simple yet widely adopted implementation in SFDA, our approach focuses on effectively controlling uncertainty during the adaptation process by refining both the negative and positive sample components.

Dispersion Control via Data Augmentation Alignment. To minimize the effect of false negative samples - points from the same class as the anchor point but misidentified as negative examples, we introduce a dispersion control term $\mathcal{L}_{\mathrm{DC}}^{-}$, which complements the conventional negative sample loss $\mathcal{L}_{\mathrm{CL}}^{-}$. This leads to the following negative uncertainty control loss:

$$\mathcal{L}_{\mathrm{UCon}}^{-} \triangleq \lambda_{\mathrm{CL}}^{-}\mathcal{L}_{\mathrm{CL}}^{-} + \lambda_{\mathrm{DC}}\mathcal{L}_{\mathrm{DC}}^{-} \triangleq \frac{1}{N_{\mathrm{T}}}\sum_{i=1}^{N_{\mathrm{T}}}\big\{\lambda_{\mathrm{CL}}^{-}\sum_{\mathbf{x}_i^{-}\in\mathcal{B}\backslash\{\mathbf{x}_i\}}\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}_i^{-};\mathbf{x}_i) - \lambda_{\mathrm{DC}}\mathbb{d}_{\boldsymbol{\theta}}\left(\mathtt{AUG}\left(\mathbf{x}_i\right),\mathbf{x}_i\right)\big\}, \tag{7}$$

where where $\mathbb{d}_{\boldsymbol{\theta}}\left(\mathtt{AUG}\left(\mathbf{x}_i\right),\mathbf{x}_i\right) = \langle f_{\mathrm{T}}(\mathbf{x}_i;\boldsymbol{\theta}), \log f_{\mathrm{T}}\left(\mathtt{AUG}\left(\mathbf{x}_i\right);\boldsymbol{\theta}\right)\rangle$ is the cosine similarity between network output of $\mathbf{x}_i$ and the log probabilities of its augmented version. $\mathcal{B}$ denotes the mini-batch, and N_{T} represents the size of the target data set. For data augmentation, we use the general augmentation pipeline proposed in self-supervised learning Chen et al. (2020). Similar to previous work (Yang et al., 2022), the decay coefficient $\lambda_{\mathrm{CL}}^{-}$ is defined as $\lambda_{\mathrm{CL}}^{-} = (1 + 10 * \frac{iter}{max\_iter})^{\beta}$, with β and λ_{DC} being hyperparameters.

Different from previous works that exclude false negative samples (Chen et al., 2022; Litrico et al., 2023) or adjust the coefficient $\lambda_{\mathrm{CL}}^{-}$ (Mitsuzumi et al., 2024a), our proposed dispersion control term intelligently utilizes data augmentation to mimic false negatives without introducing additional uncertainty. This approach implicitly reduces the variance in prediction similarity between anchor data and noisy negative samples while enhancing the model's prediction consistency.

Supervision Relaxation by Partial Label Training. As highlighted in Theorem 4.2, partial labels can help control uncertainty in positive sample predictions in SFDA. Our findings show that neighboring samples in the feature space can sufficiently provide accurate label information for initially confident target samples, but highly uncertain samples require additional processing. To handle these uncertain samples, we propose an innovative approach to select uncertain samples during adaptation by tracking the ratio between the largest and second-largest predicted probabilities. Specifically, we maintain an uncertain data bank, defined as: $\mathcal{U} = \{\mathbf{x}\in\mathcal{D}_{\mathrm{T}} : \frac{f_{\mathrm{T}}(\mathbf{x};\boldsymbol{\theta})_{(1)}}{f_{\mathrm{T}}(\mathbf{x};\boldsymbol{\theta})_{(2)}} \le \tau\}$, where $f_{\mathrm{T}}(\mathbf{x};\boldsymbol{\theta})_{(i)}$ is the i-largest predicted probabilities for $\mathbf{x}$. The threshold τ is typically set to a small value, usually between 1 and 1.5, to capture severely uncertain samples. Additionally, we store the historical TOP-K_{PL} predicted labels for each data $\mathbf{x}_i$ to construct a partial label set, denoted as $\mathcal{Y}_{\mathrm{PL},i}$, which is then used to further supervise the training of uncertain data. Aftering incorporating the partial label loss $\mathcal{L}_{\mathrm{PL}}^{+}$, the positive uncertainty control loss term $\mathcal{L}_{\mathrm{UCon}}^{+}$ is defined as:

$$\mathcal{L}_{\mathrm{UCon}}^{+} \triangleq \mathcal{L}_{\mathrm{CL}}^{+} + \lambda_{\mathrm{PL}}\mathcal{L}_{\mathrm{PL}}^{+} \tag{8}$$

$$\triangleq \frac{1}{N_{\mathrm{T}}}\sum_{i=1}^{N_{\mathrm{T}}}\big\{-\sum_{\mathbf{x}_i^{+}\in\mathcal{C}_i}\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}_i^{+};\mathbf{x}_i) + \lambda_{\mathrm{PL}}\sum_{\mathrm{y}_{k,i}\in\mathcal{Y}_{\mathrm{PL},i}}\mathbb{1}_{\{\mathbf{x}_i\in\mathcal{U}\}}\ell_{\mathrm{CE}}(\mathrm{y}_{k,i}, f_{\mathrm{T}}(\mathbf{x}_i;\boldsymbol{\theta}))\big\}, \tag{9}$$

where $\mathcal{C}_i$ is the neighbor set of $\mathbf{x}_i$, $\mathbb{1}$ is the indicator function, ℓ_{CE} is the smoothed cross-entropy loss, and λ_{PL} is a hyperparameter.

Unlike most uncertainty-based approaches in SFDA, which focus on excluding or reducing the negative impact of highly uncertain data during adaptation (Roy et al., 2022; Litrico et al., 2023), our method leverages uncertainty to extract additional label information from these data, relaxing the training process and boosting the performance.

Overall Uncertainty Control SFDA Loss. The final Uncertainty Control SFDA loss $\mathcal{L}_{\mathrm{UCon-SFDA}}$ is defined as:

$$\mathcal{L}_{\mathrm{UCon-SFDA}} = \mathcal{L}_{\mathrm{CL}} + \lambda_{\mathrm{PL}}\mathcal{L}_{\mathrm{PL}}^{+} + \lambda_{\mathrm{DC}}\mathcal{L}_{\mathrm{DC}}^{-}. \tag{10}$$

The pseudocode for the algorithm (Algorithm 1) and the complete training process can be found in the Appendix B.

Table 1: Classwise Accuracy (%) on the VisDA2017 Dataset (ResNet-101): Synthetic → Real

Method	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Per-class
3C-GAN (Li et al., 2020b)	94.8	73.4	68.8	74.8	93.1	95.4	88.6	**84.7**	89.1	84.7	83.5	48.1	81.6
SHOT (Liang et al., 2020)	94.3	88.5	80.1	57.3	93.1	94.9	80.7	80.3	91.5	89.1	86.3	58.2	82.9
A^{2}Net (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	83.3	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
CPGA (Qiu et al., 2021)	95.6	89.0	75.4	64.9	91.7	**97.5**	89.7	83.8	93.9	93.4	87.7	**69.0**	86.0
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**	47.9	86.8
CoWA-JMDS (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
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	64.7	88.0
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
SF(DA)^2 (Hwang et al., 2024)	96.8	89.3	82.9	**81.4**	96.8	95.7	90.4	81.3	95.5	93.7	88.5	64.7	88.1
I-SFDA (Mitsuzumi et al., 2024a)	97.5	**91.4**	87.9	79.4	97.2	97.2	92.2	83.0	96.4	94.2	91.1	53.0	88.4
UCon-SFDA (Ours)	**98.4**	90.7	**88.6**	80.7	**97.9**	96.9	**93.1**	83.8	**97.6**	**95.9**	92.6	59.1	**89.6**

5 Experiments

5.1 Experimental Setup

Datasets. To evaluate the proposed method, we conduct experiments on several SFDA benchmarks under three different domain shift scenarios: general SFDA, SFDA with severe label shift, and source-free partial set domain adaptation. For general SFDA, we test our method on the **Office-31** (Saenko et al., 2010), **Office-Home** (Venkateswara et al., 2017), **VisDA2017** (Peng et al., 2017), and **DomainNet-126** (Litrico et al., 2023) datasets. **VisDA2017** is a relatively large-scale classification dataset with 12 classes, consisting of 152K synthetic images and 55K real-world object images. We use the synthetic images as the source domain and the real images as the target domain. **Office-31** contains 4,652 images from three domains (Amazon, DSLR, and Webcam) across 31 categories, while **Office-Home** comprises 15,550 images from four domains (Real, Clipart, Art, and Product) with 65 classes. **DomainNet-126** is a subset of the larger DomainNet dataset that includes over 600K images across 345 categories and six domains (Clipart, Infograph, Painting, Quickdraw, Real and Sketch) (Peng et al., 2019). Following previous work (Litrico et al., 2023), we use 126 selected classes from four of these sub-domains for our experiments.

We further test on more complex SFDA tasks. For source-free domain adaptation with label shift, we employ the **VisDA-RUST** dataset, which presents a severe label imbalance in the target domain (Li et al., 2021). For source-free partial set domain adaptation, we follow the setup in Liang et al. (2020) for the **Office-Home** dataset, where only the first 24 classes are retained in the target domain.

Implementation Details. To ensure fair experimental comparisons, we use the same neural network architectures and training schemes as in previous state-of-the-art approaches (Liang et al., 2020; Yang et al., 2022; Karim et al., 2023; Hwang et al., 2024). Specifically, we adopt ResNet-50 as the backbone model for the Office-31, Office-Home, and DomainNet-126 datasets, and ResNet-101 for VisDA. We replace the original fully connected layer in ResNet with a bottleneck layer followed by batch normalization, and then add a simple linear layer with weight normalization for the classification. For adaptation training on the target domain, we use the SGD optimizer with the same learning rate scheduler as in Liang et al. (2020). For evaluation, we report the average accuracy for Office-31, Office-Home, and DomainNet-126. For VisDA2017 and VisDA-RUST, we report both per-class top-1 accuracy and the overall average. All experiments are run with three random seeds, and the average results are reported. Further implementation details, including the hyperparameter selection, can be found in Appendix B.

5.2 Overall Experimental Results

The experimental results are summarized in Tables 1- 4 and Table 7 in Appendix C.1, with the best result highlighted in bold. Our proposed method consistently outperforms all baseline methods, especially on the large-scale datasets VisDA2017 (+1.2%) and DomainNet-126 (+1.9%). For VisDA2017, a dataset with only 12 classes, conventional negative sample selection methods that treat the entire batch as negative samples often introduce significant noise and uncertainty. By incorporating the negative sample uncertainty loss, we investigate this issue and see a notable

Table 2: Classwise Accuracy (%) on the VisDA-RSUT Dataset (ResNet-101): Synthetic → Real

Method	plane	bcycle	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Per-class
Source only (He et al., 2016)	79.9	15.7	40.6	**77.2**	66.8	11.1	85.1	12.9	48.3	14.3	64.6	3.3	43.3
SHOT (Liang et al., 2020)	**86.2**	**48.1**	77.0	62.8	92.0	66.2	90.7	61.3	76.9	73.5	67.2	9.1	67.6
CoWA-JMDS (Lee et al., 2022)	63.8	32.9	69.5	59.9	93.2	95.4	92.3	69.4	85.1	68.4	64.9	32.3	68.9
NRC (Yang et al., 2021a)	**86.2**	47.6	66.7	68.1	94.7	76.6	93.7	63.6	87.3	89.0	83.6	20.0	73.1
AaD (Yang et al., 2022)	73.9	33.3	56.6	71.4	90.1	97.0	91.9	70.8	88.1	87.2	81.2	**39.4**	73.4
SF(DA)^2 (Hwang et al., 2024)	79.0	43.3	73.6	74.7	92.8	**98.3**	93.4	79.1	90.1	87.5	81.1	34.2	77.3
UCon-SFDA (Ours)	84.1	37.1	**87.4**	70.6	**95.4**	92.9	**94.4**	**83.0**	**93.7**	**92.0**	**86.7**	35.3	**79.4**

Table 3: Classification Accuracy (%) on the Office-Home Dataset (ResNet-50) Under Source-Free Partial-Set Domain Adaptation

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., 2020)	64.8	85.2	92.7	76.3	77.6	**88.8**	79.7	64.3	89.5	80.6	66.4	85.8	79.3
AaD (Yang et al., 2022)	**67.0**	83.5	**93.1**	**80.5**	76.0	87.6	78.1	65.6	**90.2**	**83.5**	64.3	87.3	79.7
UCon-SFDA (Ours)	65.6	**87.8**	91.0	78.6	**79.3**	87.6	**80.2**	**65.9**	87.3	83.2	**69.1**	**88.7**	**80.3**

performance boost. Similarly, our method excels in more challenging tasks, such as Ar → Cl and Pr → Cl on Office-Home, and it consistently performs well across nearly all tasks on DomainNet-126. Additional experimental results and analyses, including self-prediction accuracy, data augmentation consistency, variance control effect, hyperparameter sensitivity, performance under various similarity measures utilized in dispersion control term and complexity analyses, are provided in Appendix C.

In more complex scenarios like VisDA-RUST (with severe label imbalance), we observe a performance gain of +2.1%, while for the partial set Office-Home setup, our method shows a +0.6% improvement. These results further confirm the robustness and generality of our proposed method, particularly in handling highly imbalanced target domain data and challenging source-free domain adaptation tasks.

5.3 Analysis

Ablation Study. To evaluate the effectiveness and necessity of each component proposed in our algorithm, we conduct an ablation study across four datasets. The results, shown in Table 5, demonstrates that both partial label supervision training and dispersion control can boost the performance of the baseline approach ($\mathcal{L}_{\text{CL}}$). While $\mathcal{L}^{+}_{\text{PL}}$ can better handle severe label shift scenarios, as seen in the VisDA-RUST dataset, $\mathcal{L}^{-}_{\text{DC}}$ performs better on more difficult tasks. Notably, adding the dispersion control term alone improves or matches the performance of most negative sample denoising and uncertainty-based methods, such as those from Roy et al. (2022); Litrico et al. (2023); Chen et al. (2022); Mitsuzumi et al. (2024a), without requiring any additional networks. Combining both positive and negative uncertainty control can boost each other and enhance the performance.

Negative Sampling Dispersion Control. To further evaluate the effect of the dispersion control by $\mathcal{L}^{-}_{\text{DC}}$, we calculate the variance in prediction similarity between anchor-true-negative pairs during adaptation. Figure 3c illustrates that introducing $\mathcal{L}^{-}_{\text{DC}}$ succesfully reduces this variance. Further more, the SF(DA)^2 method (Hwang et al., 2024) approaches the problem from a graph-based perspective

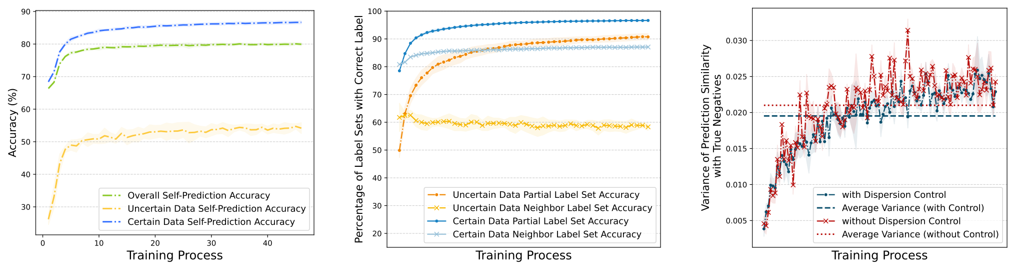
Table 4: Classification Accuracy (%) on Office-31 (left) and DomainNet-126 (right) using ResNet-50

Method	A → D	A → W	D → W	W → D	D → A	W → A	Avg.
SHOT (Liang et al., 2020)	94.0	90.1	98.4	99.9	74.7	74.3	88.6
3C-GAN (Li et al., 2020b)	92.7	93.7	98.5	99.8	75.3	77.8	89.6
A^{2}Net (Xia et al., 2021)	94.5	94.0	**99.2**	**100.0**	76.7	76.1	90.1
NRC (Yang et al., 2021a)	96.0	90.8	99.0	**100.0**	75.3	75.0	89.4
CPGA (Qiu et al., 2021)	94.4	94.1	98.4	99.8	76.0	76.6	89.9
CoWA-JMDS (Lee et al., 2022)	94.4	95.2	98.5	99.8	76.2	77.6	90.3
AaD (Yang et al., 2022)	**96.4**	92.1	99.1	**100.0**	75.0	76.5	89.9
C-SFDA (Karim et al., 2023)	96.2	93.9	98.8	99.7	**77.3**	**77.9**	90.5
I-SFDA (Mitsuzumi et al., 2024a)	95.3	94.2	98.3	99.9	76.4	77.5	90.3
UCon-SFDA (Ours)	94.8	**95.4**	98.9	**100.0**	77.1	77.1	**90.6**

Method	S→P	C→S	P→C	P→R	R→S	R→C	R→P	Avg.
Source only (He et al., 2016)	50.1	46.9	53.0	75.0	46.3	55.5	62.7	55.6
TENT (Wang et al., 2020)	52.4	48.5	57.9	67.0	54.0	58.5	65.7	57.7
DivideMix (Li et al., 2020a)	64.3	61.3	67.7	77.3	62.4	68.1	69.5	67.2
SHOT (Liang et al., 2020)	66.1	60.1	66.9	80.8	59.9	67.7	68.4	67.1
NRC (Yang et al., 2021a)	65.7	58.6	64.5	82.3	58.4	65.2	68.2	66.1
AaD (Yang et al., 2022)	65.4	54.2	59.8	81.8	54.6	60.3	68.5	63.5
AdaContrast (Chen et al., 2022)	65.9	58.0	68.6	80.5	61.5	70.2	69.8	67.8
GPUE (Litrico et al., 2023)	67.5	64.0	68.8	76.5	**65.7**	74.2	70.4	69.6
SF(DA)^2 (Hwang et al., 2024)	67.7	59.6	67.8	**83.5**	60.2	68.8	70.5	68.3
UCon-SFDA (Ours)	**68.1**	**66.5**	**69.3**	81.0	64.3	**75.2**	**71.1**	**71.5**

Table 5: Ablation Study Results across Different Datasets and Tasks

Method	VisDA2017	VisDA-RUST	DomainNet-126			OfficeHome		
	Sync → Real	Sync → Real	P → R	R → P	Avg.	Ar → Cl	Pr → Cl	Avg.
$\mathcal{L}_{\mathrm{CL}}$	87.6	75.5	78.9	67.8	66.9	58.6	57.9	72.6
$\mathcal{L}_{\mathrm{CL}} + \mathcal{L}^{-}_{\mathrm{DC}}$	89.0	78.9	80.2	70.3	69.8	61.2	59.7	73.3
$\mathcal{L}_{\mathrm{CL}} + \mathcal{L}^{+}_{\mathrm{PL}}$	88.1	79.1	80.8	69.5	68.8	60.2	59.3	73.1
$\mathcal{L}_{\mathrm{UCon-SFDA}}$	89.6	79.4	81.0	71.1	71.5	61.5	62.2	73.6



(a) Self-Prediction Accuracies (b) Partial Label Set Quality (c) Similarity Dispersion Control

Figure 3: *(a) Self-Prediction Accuracies across data with varying levels of predictive uncertainty on Office-Home (Ar → Pr). (b) Comparison of the quality of partial label set and neighbor label set across different uncertainty levels. (c) Comparison of prediction similarity variances between anchor-true negative sample pairs with and without the dispersion control term $\mathcal{L}^{-}_{\mathrm{DC}}$ on Office-Home (Ar → Cl).*

and introduces a quadratic regularized term on the predicted probability similarity of anchor-negative pairs. It is equivalent to directly minimizing the variance. Our experimental results also demonstrates the effectiveness of our data augmentation-based dispersion control.

Positive Supervision Uncertainty Relaxation. As shown in Figure 3a, the top-1 self-predicted label is more accurate for certain data points (blue dot line in Figure 3a) than uncertain ones (yellow dot line), which indicates that uncertain data require additional supervision during adaptation. To further validate the proposed partial label supervision on these uncertain target data, we define a neighbor label set that contains the neighbors' self-predicted top-1 label. We compare the label information provided by this neighbor label set against our proposed partial label set. By comparing the two lines representing the accuracy of the neighbor label sets marked with 'x' in Figure 3b, we can easily observe that for uncertain data, neighbor label set becomes increasingly unstable as training progresses, with accuracy sometimes even decreasing. This highlights why we choose not to rely on neighbor labels in our algorithm design. Instead, we use the sample's own TOP-K_{PL} predictions to form a partial label set. A closer look at the difference between the two blue lines and the two yellow lines in Figure 3b reveals that the label gain from the partial label set is much greater for uncertain data than for certain data. Interestingly, the accuracy of the neighbor's labels is consistently higher than the overall accuracy of the model's self-prediction, which explains why we only apply relaxed supervision through partial label loss for uncertain data.

6 Conclusion

In this paper, we thoroughly analyze two types of uncertainty in SFDA arising from the use of positive and negative samples. By examining the uncertainty in the negative sample distribution during training, we construct an outlier-robust worst-case risk and derive an informative upper bound for it. This analysis not only explains why current contrastive learning methods significantly enhance SFDA performance but also leads to the design of an augmentation-based dispersion control approach to mitigate the uncertainty introduced by noisy negative samples. Furthermore, by investigating the prediction uncertainty of positive examples, we identify a partial label set as the optimal solution for the target data. This revelation uncovers previously overlooked uncertain information in existing algorithms and motivates us to propose novel criteria for distinguishing uncertain data, thereby using partial labels to relax the supervision from positive examples.

REFERENCES

Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28 (1):131–142, 1966.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

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.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3941–3950, 2020.

John Duchi. Lecture notes for statistics 311/electrical engineering 377. *URL: https://stanford. edu/class/stats311/Lectures/full_notes. pdf. Last visited on*, 2:23, 2016.

John Duchi. Information theory and statistics. *Lecture Notes for Statistics*, 311, 2019.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Rui Gao. Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 71(6):2291–2306, 2023.

Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 72(3):1177–1191, 2024.

Lars Peter Hansen and Thomas J Sargent. *Robustness*. Princeton university press, 2008.

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.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Uiwon Hwang, Jonghyun Lee, Juhyeon Shin, and Sungroh Yoon. Sf(da)^2: Source-free domain adaptation through the lens of data augmentation. In *International Conference on Learning Representations*, 2024.

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.

Henry Lam. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275, 2016.

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, 2022.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020a.

Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9641–9650, 2020b.

Xinhao Li, Jingjing Li, Lei Zhu, Guoqing Wang, and Zi Huang. Imbalanced source-free domain adaptation. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3330–3339, 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 *Proceedings of the International Conference on Machine Learning*, pp. 6028–6039, 2020.

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.

David G Luenberger and Yinyu Ye. *Linear and Nonlinear Programming*, volume 2. Springer, 1984.

Yu Mitsuzumi, Akisato Kimura, and Hisashi Kashima. Understanding and improving source-free domain adaptation from a theoretical perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28515–28524, 2024a.

Yu Mitsuzumi, Akisato Kimura, and Hisashi Kashima. Understanding and improving source-free domain adaptation from a theoretical perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28515–28524, 2024b.

Sloan Nietert, Ziv Goldfeld, and Soroosh Shafiee. Outlier-robust wasserstein dro. *Advances in Neural Information Processing Systems*, 36, 2024a.

Sloan Nietert, Ziv Goldfeld, and Soroosh Shafiee. Robust distribution learning with local and global adversarial corruptions (extended abstract). In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247, pp. 4007–4008, 2024b.

Jiangbo Pei, Zhuqing Jiang, Aidong Men, Liang Chen, Yang Liu, and Qingchao Chen. Uncertainty-induced transferability representation for source-free unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 32:2033–2048, 2023.

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.

Zhen Qiu, Yifan Zhang, Hongbin Lin, Shuaicheng Niu, Yanxia Liu, Qing Du, and Mingkui Tan. Source-free domain adaptation via avatar prototype generation and adaptation. *arXiv preprint arXiv:2106.15326*, 2021.

Subhankar Roy, Martin Trapp, Andrea Pilzer, Juho Kannala, Nicu Sebe, Elisa Ricci, and Arno Solin. Uncertainty-guided source-free domain adaptation. In *European conference on computer vision*, pp. 537–555, 2022.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision*, pp. 213–226, 2010.

Kuniaki Saito, Donghyun Kim, Piotr Teterwak, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9184–9193, 2021.

C Shalizi. Almost none of stochastic processes, 2006.

Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 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.

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.

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.

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*, pp. 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, Shangling Jui, Joost van de Weijer, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 5802–5815, 2022.

Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. Doro: Distributional and outlier robust optimization. In *International Conference on Machine Learning*, pp. 12345–12355. PMLR, 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. *Advances in Neural Information Processing Systems*, 35:5137–5149, 2022.

A TECHNICAL DETAILS

A.1 NOTATION TABLE

The notation table provides a summary for the key notations used throughout the paper, with the symbols, descriptions, and the first appearance place included in the first, second, an the third columns, respectively.

Notations	**Descriptions**	First appearance
$\mathcal{X} \subset \mathbb{R}^d$	d-dimensional input space	Section 3
$\mathcal{Y} = [K]$	label space for K-classification	Section 3
$P^{\rm S}_{\mathbf{xy}}$; $\mathcal{D}_{\rm S}$	underlying distribution over $\mathcal{X} \times \mathcal{Y}$ related to source domain unavailable source domain data $\mathcal{D}_{\rm S} \triangleq \{\mathbf{x}_i^{\rm S}, \mathrm{y}_i^{\rm S}\}_{i=1}^{N_{\rm S}}$	Section 3
$P^{\rm T}_{\mathbf{xy}}$; $\mathcal{D}_{\rm T}$	underlying distribution over $\mathcal{X} \times \mathcal{Y}$ related to target domain unlabeled target domain data $\mathcal{D}_{\rm T} \triangleq \{\mathbf{x}_i^{\rm T}\}_{i=1}^{N_{\rm T}}$	Section 3
$f_{\rm S}(\mathbf{x};\boldsymbol{\theta})/f_{\rm T}(\mathbf{x};\boldsymbol{\theta})/f(\mathbf{x};\boldsymbol{\theta}):$ $\mathcal{X} \mapsto \Delta^{K-1}$	predicted probabilities of source/target/general model	Section 3
$h_{\rm S}(\mathbf{x};\boldsymbol{\theta})/h_{\rm T}(\mathbf{x};\boldsymbol{\theta})/h(\mathbf{x};\boldsymbol{\theta}):$ $\mathcal{X} \mapsto \mathcal{Y}$	source/target/general classifier: $= \arg\max_{j\in[K]} f_{\rm S}(\mathbf{x};\boldsymbol{\theta})[j]/f_{\rm T}(\mathbf{x};\boldsymbol{\theta})[j]/f(\mathbf{x};\boldsymbol{\theta})[j]$	Section 3
$\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}';\mathbf{x})$	similarity between $\mathbf{x}'$ and $\mathbf{x}$ e.g., $\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}';\mathbf{x}) =< f(\mathbf{x}';\boldsymbol{\theta}), f(\mathbf{x};\boldsymbol{\theta}) >$	Section 4.1, Eq. (1)
$P^{\rm T}_{\mathbf{x}}$ (empirical: $\widehat{P}_{\mathbf{x}}$)	distribution of input $\mathbf{X}$ (target)	Section 3
$P^{+}(\cdot;\mathbf{x})$, or simply P^{+} (empirical: $\widehat{P}^{+}$)	conditional distribution for positive sample over $\mathcal{X}$, given $\mathbf{x}$	Section 4.1, Eq. (1)
$P^{-}(\cdot;\mathbf{x})$, or simply P^{-} (empirical: $\widehat{P}^{-}$)	conditional distribution for negative sample over $\mathcal{X}$, given $\mathbf{x}$	Section 4.1, Eq. (1)
$\mathcal{L}^{+}_{\rm CL}$ / $\mathcal{L}^{-}_{\rm CL}$	positive/negative contrastive loss	Section 4.3, Remark 4.4
$\mathcal{L}^{+}_{\rm PL}$ / $\mathcal{L}^{-}_{\rm DC}$	partial label/dispersion control loss	Section 4.3, Remark 4.4
$\mathcal{L}^{+}_{\rm UCon}$ / $\mathcal{L}^{-}_{\rm UCon}$	overall positive/negative uncertainty control loss	Section 4.4, Eq. (8)
$\mathcal{L}_{\rm UCon-SFDA}$	uncertainty control source-free domain adaptation loss	Section 4.4, Eq. (10)
$\lambda_{\rm PL}$ / $\lambda_{\rm DC}$ / $\lambda^{-}_{\rm CL}$	partial label/dispersion control/negative contrastive loss coefficient	Section 4.4, Eq. (8) / (7) / (6)
κ	number of neighbors for each anchor point	Section 4.1
$K_{\rm PL}$	update number for partial label set	Section 4.4 (Page 7)
τ	uncertain sample selection ratio	Section 4.4 (Page 7)
β	decay exponent of negative contrastive loss	Section 4.4 (Page 7)
$\mathscr{Z}$ / $\mathscr{F}$ / $\mathscr{Y}_{\rm PL}$ / $\mathscr{U}$	feature/predicted probabilities/ partial label set/uncertainty sample bank	Appendix B, Algorithm 1
$\texttt{AUG}(\mathbf{x})$	data augmentation of input sample $\mathbf{x}$	Section 4.2, Remark 4.2

A.2 PRELIMINARIES ON DISCREPANCY METRICS AND LINEAR PROGRAMMING

We begin by presenting the definitions and some optimization results of the p-Wasserstein distance and φ-divergence, which are potential choices for the discrepancy metric d in (3), and will be used in the proof of Theorem 4.1.

Definition A.1 (p-Wasserstein distance (Blanchet & Murthy, 2019))**.** For a Polish space Ω (i.e., a complete separable metric space) endowed with a metric $c : \Omega \times \Omega \to \mathbb{R}_{\geq 0}$, let $\mathcal{P}(\Omega)$ represent the set of all Borel probability measures on Ω, where $\mathbb{R}_{\geq 0}$ represents the set of all nonnegative real values. For $p \geq 1$, let $\mathcal{P}_p(\Omega)$ stand for the subset of $\mathcal{P}(\Omega)$ with finite pth moments. Then, for $P_1, P_2 \in \mathcal{P}_p(\Omega)$, the Wasserstein distance of order p is defined as

$$W_p(P_1, P_2) \triangleq \inf_{\Pi \in \mathrm{Cpl}(P_1, P_2)} \left[\mathbb{E}_{(S_1,S_2)\sim\Pi}\left\{c^p(S_1, S_2)\right\}\right]^{1/p},$$

where $\mathrm{Cpl}(P_1, P_2)$, sometimes called the coupling set of P_1 and P_2, comprises all probability measures on the product space $\Omega \times \Omega$ such that their marginal measures are $P_1(\cdot)$ and $P_2(\cdot)$. Here, $c^p(\cdot,\cdot)$ represents $\{c(\cdot,\cdot)\}^p$.

Definition A.2 (φ-divergence (Ali & Silvey, 1966; Duchi, 2019))**.** Let P and Q be probability distributions on a measure space $(\Omega, \mathcal{G})$, and let $\varphi : \mathbb{R}_+ \longrightarrow \mathbb{R}$ be a convex function satisfying

$\varphi(1) = 0$ and $\varphi(t) = +\infty$ for $t < 0$. Without loss of generality, assume that P and Q are absolutely continuous with respect to the base measure μ. The φ-divergence between P and Q is then defined as

$$D_\varphi(P\|Q) := \int_\Omega q(x)\varphi\left(\frac{p(x)}{q(x)}\right) d\mu(x) + f'(\infty)P\{q=0\},$$

where p and q are the densities of P and Q with respect to the measure μ, respectively, and $\varphi'(\infty)$ represents $\lim_{x\to\infty} \varphi(t)/t$.

Example A.1 (Duchi, 2019, Chapter 2.2)**.** By taking different φ functions, we provide some popular examples of φ-divergences.

- Kullback-Leibler (KL) divergence: taking $\varphi(t) = t\log t$ gives $D_\varphi(P\|Q) \triangleq D_{\mathrm{KL}}(P\|Q) = \int p\log(p/q)d\mu$.
- The total variation distance: taking $\varphi(t) = \frac{1}{2}|t-1|$ yields $D_\varphi(P\|Q) \triangleq \|P-Q\|_{\mathrm{TV}} = \frac{1}{2}\int\left|\frac{p}{q}-1\right|qd\mu = \sup_{A\subset\Omega}|P(A)-Q(A)|$.
- The Hellinger distance: taking $\varphi(t) = (\sqrt{t}-1)^2 = t - 2\sqrt{t} + 1$ leads to the squared Hellinger distance $D_\varphi(P\|Q) \triangleq H^2(P\|Q) = \int(\sqrt{p}-\sqrt{q})^2 d\mu$.
- The χ^2-divergence: taking $\varphi(t) = (t-1)^2$ produces the χ^2-divergence $D_\varphi(P\|Q) \triangleq \chi^2(P\|Q) = \int(\frac{p}{q}-1)^2 d\mu$.

Lemma 1 (Strong duality for robust risk based on p-Wasserstein distance (Gao et al., 2024, Lemma EC.1))**.** *Consider the p-Wasserstein distance $W_p(\cdot,\cdot)$ with $p \in [1,\infty)$ defined in Definition A.1. Given a upper semi-continuous loss function $\hbar : \Omega \to \mathbb{R}$, a nominal distribution $P \in \mathcal{P}_p(\Omega)$, and a radius $\delta > 0$, the corresponding robust risk based on the p-Wasserstein distance $W_p(\cdot,\cdot)$ is*

$$\mathfrak{v}_{\mathrm{P}} \triangleq \sup_{Q\in\mathcal{P}(\Omega)} \left[\mathbb{E}_{Z\sim Q}\big\{\hbar(Z)\big\} : W_p(P,Q) \le \delta\right].$$

The dual problem is defined as

$$\mathfrak{v}_{\mathrm{D}} \triangleq \min_{\gamma\ge 0}\left\{\gamma\delta^p + \mathbb{E}_{Z\sim P}\left[\sup_{z'\in\Omega}\left\{\hbar(z') - \gamma c^p(z',Z)\right\}\right]\right\}.$$

Then, $\mathfrak{v}_{\mathrm{P}} = \mathfrak{v}_{\mathrm{D}}$.

Lemma 2 (Strong duality for robust risk based on φ-divergence (Duchi & Namkoong, 2021, Proposition 1; Shapiro, 2017, Section 3.2))**.** *Consider the φ-divergence $D_\varphi(\cdot\|\cdot)$ defined in Definition A.2. Given a loss function $\hbar : \Omega \to \mathbb{R}$, a nominal distribution P on the measure space $(\Omega, \mathcal{G})$, and a radius $\delta > 0$, the corresponding robust risk based on the φ-divergence $D_\varphi(\cdot\|\cdot)$ is*

$$\mathfrak{v}_{\mathrm{P}} \triangleq \sup_{Q\ll P}\left[\mathbb{E}_{Z\sim Q}\big\{\hbar(Z)\big\} : D_\varphi(Q||P) \le \delta\right].$$

The dual problem is defined as

$$\mathfrak{v}_{\mathrm{D}} \triangleq \inf_{\gamma\ge 0,\eta\in\mathbb{R}}\left\{\mathbb{E}_P\left[\gamma\varphi^*\left\{\frac{\hbar(Z)-\eta}{\gamma}\right\}\right] + \gamma\delta + \eta\right\},$$

where $\varphi^*(t) = \sup_s\{ts - \varphi(s)\}$ for any $t \in \mathbb{R}$ is the Fenchel conjugate. Then, $\mathfrak{v}_{\mathrm{P}} = \mathfrak{v}_{\mathrm{D}}$. Moreover, if the supremum in $\mathfrak{v}_{\mathrm{P}}$ is finite, then there are finite $\gamma \ge 0$ and $\eta \in \mathbb{R}$ attaining the infimum in $\mathfrak{v}_{\mathrm{D}}$.

Lemma 3 (Hansen & Sargent, 2008, Proposition 1.4.2)**.** *Let $(\Omega, \mathcal{G}, \mu)$ represent a σ-finite measure space, where Ω is a set, $\mathcal{G}$ is the σ-algebra of subsets of Ω, and μ is the associated measure. $\hbar : \Omega \to \mathbb{R}$ is a bounded measurable function. The following conclusions hold.*

(i) We have the variational formula

$$-\log\int_\Omega \exp\{-\hbar(\omega)\}d\mu(\omega) = \inf_{\nu\in\mathcal{P}(\Omega)}\left\{D_{KL}(\nu\|\mu) + \int_\Omega \hbar(\omega)d\nu(\omega)\right\}$$

(ii) Let ν^* denote the probability measure on Ω which is absolutely continuous with respect to μ and satisfies

$$\frac{d\nu^*}{d\mu}(\omega) \triangleq \frac{\exp\{-\hbar(\omega)\}}{\int_\Omega \exp\{-\hbar(\omega)\}d\mu(\omega)} \text{ for } \omega \in \Omega.$$

Then the infimum in the variational formula above is attained uniquely at ν^*.

We next introduce the concept of linear programming and some result on extreme points, which will be used in the proof of Theorem 4.2.

A *linear program* (LP) is an optimization problem of the form

$$\begin{aligned} \max_{\mathbf{x}\in\mathbb{R}^n} \ & \mathbf{c}^\top \mathbf{x} \\ s.t. \ & \mathbf{A}\mathbf{x} \le \mathbf{b} \\ & \mathbf{x} \ge 0, \end{aligned} \tag{11}$$

where $\mathbf{c} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^m$ are given, and $\mathbf{A}$ is a specified $m \times n$ matrix. Here, "$\le$" represents elementwise inequality for vectors. The expression $\mathbf{c}^\top \mathbf{x}$ is called the *objective function*, and the set $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \le \mathbf{b}, \mathbf{x} \ge 0\}$ defines the *feasible region* of the linear program. By introducing slack variables, any linear program can be converted to the following *standard form*:

$$\begin{aligned} \max_{\mathbf{x}\in\mathbb{R}^n} \ & \mathbf{c}^\top \mathbf{x} \\ s.t. \ & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \ge 0. \end{aligned} \tag{12}$$

Definition A.3 (Luenberger & Ye, 1984, Chapter 2)**.** A point $\mathbf{z}$ in a convex set Θ is called an *extreme point* of Θ if there do not exist two distinct points $\mathbf{z}', \mathbf{z}'' \in \Theta$ and a scalar ν with $0 < \nu < 1$ such that $\mathbf{z} = \nu\mathbf{z}' + (1-\nu)\mathbf{z}''$.

Lemma 4 (Luenberger & Ye, 1984, Chapter 2)**.** *If a linear programming problem has a finite optimal solution (i.e., a feasible solution that optimizes the objective function), then there is a finite optimal solution that is an extreme point of the constraint set.*

A.3 Proof of Theorem 4.1

Before presenting and proving the formal version of Theorem 4.1, we first examine form of the robust risk given in (2) when different choices of the discrepancy metric $\mathscr{d}$ in (3). Proof techniques in Duchi & Namkoong (2021); Zhai et al. (2021); Gao (2023); Gao et al. (2024); Lam (2016) are used.

Lemma 5. *Suppose that $\mathcal{S}_{\boldsymbol\theta}(\mathbf{X}^-; \mathbf{x}^-)$. For different choices of the discrepancy metric $\mathscr{d}$ in (3), we have the following results on the robust risk $\mathcal{R}^-_{\mathbf{x}}(\boldsymbol\theta; P^-, \delta)$ given in (2).*

(i) If $\mathscr{d}$ is the χ^2-divergence and $\delta \le \mathbb{V}_{P^-}\{\mathcal{S}_{\boldsymbol\theta}(\mathbf{X}^-; \mathbf{x})\}/\left[\mathbb{E}_{P^-}\{\mathcal{S}_{\boldsymbol\theta}(\mathbf{X}^-; \mathbf{x})\}\right]^2$, then

$$\mathcal{R}^-_{\mathbf{x}}(\boldsymbol\theta; P^-, \delta) = \mathbb{E}_{P^-}\{\mathcal{S}_{\boldsymbol\theta}(\mathbf{X}^-; \mathbf{x})\} + \sqrt{\delta\mathbb{V}_{P^-}\{\mathcal{S}_{\boldsymbol\theta}(\mathbf{X}^-; \mathbf{x})\}}.$$

(ii) If $\mathscr{d}$ is the KL-divergence, then for a small enough δ,

$$\mathcal{R}^-_{\mathbf{x}}(\boldsymbol\theta; P^-, \delta) = \mathbb{E}_{P^-}\{\mathcal{S}_{\boldsymbol\theta}(\mathbf{X}^-; \mathbf{x})\} + \sqrt{2\delta\mathbb{V}_{P^-}\{\mathcal{S}_{\boldsymbol\theta}(\mathbf{X}^-; \mathbf{x})\}} + \mathcal{O}(\delta).$$

(iii) Suppose $\mathscr{d}$ is the p-Wasserstein distance with $p \in [1, +\infty)$ and the cost function $c(\cdot,\cdot)$ in Definition A.1 is chosen as a norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$. Assume the following smoothness condition are true.

a. For any $\widetilde{\mathbf{x}}^-, \mathbf{x}^-, \mathbf{x} \in \mathcal{X}$, $\exists \mathscr{M}_1, \mathscr{M}_2 > 0$ and $\zeta \in [1, p]$, such that $\|\nabla\mathcal{S}_{\boldsymbol\theta}(\widetilde{\mathbf{x}}^-; \mathbf{x}) - \nabla\mathcal{S}_{\boldsymbol\theta}(\mathbf{x}^-; \mathbf{x})\|_* \le \mathscr{M}_1 + \mathscr{M}_2\|\widetilde{\mathbf{x}}^- - \mathbf{x}^-\|^{\zeta-1}$.

b. There exists $\eta_0 > 0$ and $\mathscr{M}_3 > 0$, such that for any $\widetilde{\mathbf{x}}^-, \mathbf{x}^-, \mathbf{x} \in \mathcal{X}$, if $\|\widetilde{\mathbf{x}}^- - \mathbf{x}^-\| \le \eta_0$, then $\|\nabla\mathcal{S}_{\boldsymbol\theta}(\widetilde{\mathbf{x}}^-; \mathbf{x}) - \nabla\mathcal{S}_{\boldsymbol\theta}(\mathbf{x}^-; \mathbf{x})\|_* \le \mathscr{M}_3\|\widetilde{\mathbf{x}}^- - \mathbf{x}^-\|$.

Let q denote the Hölder number of p, that is $\frac{1}{p}+\frac{1}{q}=1$. Then

$$\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta};P^{-},\delta)\leq\mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}+\delta\left\{\mathbb{E}_{P^{-}}\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_*^q\right\}^{1/q}+\mathcal{O}(\delta^{2\wedge p}).$$

Proof. We explore the upper bound form of $\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta};P^{-},\delta)$ under various choices of the discrepancy metric d in (3).

Case 1: χ^2-divergence. For χ^2-divergence, we have $\varphi(t)=(t-1)^2$ for $t\geq 0$ and $\varphi(t)=+\infty$ for $t<0$ by Example A.1. The Fenchel conjugate of φ is given as:

$$\begin{aligned}\varphi^*(t)&=\sup_{s\in\mathbb{R}}\{ts-\varphi(s)\}=\sup_{s\geq 0}\left\{ts-(s-1)^2\right\}=\sup_{s\geq 0}\left\{-\left(s-\frac{t+2}{2}\right)^2+\frac{t^2}{4}+t\right\}\\&=\begin{cases}\dfrac{t^2}{4}+t, \text{ for } t\geq -2\\-1, \text{ for } t<-2\end{cases}=\quad\frac{1}{4}\{(t+2)_+\}^2-1.\end{aligned}\tag{13}$$

Step (i): Upper bound on the primal problem.

If the discrepancy metric d in (3) is chosen as the χ^2-divergence, then the robust risk $\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta};P^{-},\delta)$ is expressed as

$$\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta};P^{-},\delta)=\sup_{Q^{-}\ll P^{-}}\left[\mathbb{E}_{Q^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}:\chi^2(Q^{-}\|P^{-})\leq\delta\right].\tag{14}$$

The expectation $\mathbb{E}_{Q^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}$ in $\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta};P^{-},\delta)$ can be expressed as:

$$\begin{aligned}\mathbb{E}_{Q^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}&=\mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\frac{dQ_{-}}{dP_{-}}\right\}\\&=\mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}+\mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\left(\frac{dQ_{-}}{dP_{-}}-1\right)\right\}\\&=\mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}+\mathbb{E}_{P^{-}}\left\{\left[\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}\right]\left(\frac{dQ_{-}}{dP_{-}}-1\right)\right\},\end{aligned}$$

where the first inequality holds via a change of measure and the fact that $Q^{-}\ll P^{-}$, $\frac{dQ_{-}}{dP_{-}}$ denotes the Radon–Nikodym derivative, and the last equality is true since $\mathbb{E}_{P^{-}}\left(\frac{dQ_{-}}{dP_{-}}-1\right)=0$. By Cauchy-Schwarz inequality, we further obtain that

$$\begin{aligned}&\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta};P^{-},\delta)-\mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}\\&=\sqrt{\left\{\mathbb{E}_{P^{-}}\left[\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}\right]^2\right\}\cdot\left\{\mathbb{E}_{P^{-}}\left(\frac{dQ_{-}}{dP_{-}}-1\right)^2\right\}}\\&=\sqrt{\left\{\mathbb{E}_{P^{-}}\left[\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}\right]^2\right\}\cdot\chi^2(Q^{-}\|P^{-})}\\&\leq\sqrt{\left\{\mathbb{E}_{P^{-}}\left[\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}\right]^2\right\}\cdot\delta},\end{aligned}$$

where the second equality holds by the definition of χ^2-divergence given in Example A.1, and the inequality in the last step is due to the constraint in (14). Therefore, by (14), we obtain that

$$\begin{aligned}\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta};P^{-},\delta)&\leq\mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}+\sqrt{\left\{\mathbb{E}_{P^{-}}\left[\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}\right]^2\right\}\cdot\delta}\\&\triangleq\boldsymbol{\mu}+\sqrt{\delta\boldsymbol{V}},\end{aligned}\tag{15}$$

where $\boldsymbol{\mu}\triangleq\mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}$ and $\boldsymbol{V}\triangleq\mathbb{E}_{P^{-}}\left[\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}\right]^2$.

Step (ii): Attaining the equality in the upper bound using duality.

Next, we prove the equality in the upper bound in (15) can be achieved by leveraging the strong duality result of the φ-divergence based robust risk. Specifically, according to Lemma 2 and (13),

$$\begin{aligned}\mathcal{R}_{\mathbf{x}}^{-}(\boldsymbol{\theta}; P^{-}, \delta) &= \inf_{\gamma\geq 0, \eta\in\mathbb{R}} \left\{\mathbb{E}_P\left[\gamma\varphi^*\left\{\frac{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\eta}{\gamma}\right\}\right]+\gamma\delta+\eta\right\}\\ &= \inf_{\gamma\geq 0, \eta\in\mathbb{R}} \left\{\mathbb{E}_P\left[\gamma\cdot\frac{1}{4}\left\{\frac{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\eta}{\gamma}+2\right\}_+^2-\gamma\right]+\gamma\delta+\eta\right\}\\ &= \inf_{\gamma\geq 0, \eta\in\mathbb{R}} \left[\frac{1}{4\gamma}\mathbb{E}_P\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\eta+2\gamma\right\}_+^2-\gamma+\gamma\delta+\eta\right]\\ &= \inf_{\gamma\geq 0, \widetilde{\eta}\in\mathbb{R}} \left[\frac{1}{4\gamma}\mathbb{E}_P\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\widetilde{\eta}\right\}_+^2+(1+\delta)\gamma+\widetilde{\eta}\right],\end{aligned}$$

where the last euality holds by taking $\widetilde{\eta} \triangleq \eta - 2\gamma$. By taking derivatives with respect to γ, we obtain that the optimal γ to infimize the preceding expression is given as below:

$$\gamma^* = \sqrt{\frac{\mathbb{E}_P\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\widetilde{\eta}\right\}_+^2}{4(1+\delta)}}.$$

By substituting into the preceding expression, we further obtain that

$$\mathcal{R}_{\mathbf{x}}^{-}(\boldsymbol{\theta}; P^{-}, \delta) = \inf_{\widetilde{\eta}\in\mathbb{R}}\left[\sqrt{(1+\delta)\mathbb{E}_P\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\widetilde{\eta}\right\}_+^2}+\widetilde{\eta}\right]. \tag{16}$$

Let $g(\widetilde{\eta}) \triangleq \sqrt{(1+\delta)\mathbb{E}_P\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\widetilde{\eta}\right\}_+^2}+\widetilde{\eta}$. By taking $\widetilde{\eta}^* = \boldsymbol{\mu} - \sqrt{\frac{\boldsymbol{V}}{\delta}}$, where $\boldsymbol{\mu}$ and $\boldsymbol{V}$ are defined after (15), we obtain that

$$\begin{aligned}g(\widetilde{\eta}^*) &= \sqrt{(1+\delta)\mathbb{E}_P\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\widetilde{\eta}^*\right\}_+^2}+\widetilde{\eta}^*\\ &= \sqrt{(1+\delta)\mathbb{E}_P\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\widetilde{\eta}^*\right\}^2}+\widetilde{\eta}^*\\ &= \sqrt{(1+\delta)\mathbb{E}_P\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\boldsymbol{\mu}+\sqrt{\frac{\boldsymbol{V}}{\delta}}\right\}^2}+\boldsymbol{\mu}-\sqrt{\frac{\boldsymbol{V}}{\delta}}\\ &= \sqrt{(1+\delta)\left[\mathbb{E}_P\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\boldsymbol{\mu}\right\}^2+\frac{\boldsymbol{V}}{\delta}+2\sqrt{\frac{\boldsymbol{V}}{\delta}}\mathbb{E}_P\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\boldsymbol{\mu}\right\}\right]}+\boldsymbol{\mu}-\sqrt{\frac{\boldsymbol{V}}{\delta}}\\ &= \sqrt{(1+\delta)\left(\boldsymbol{V}+\frac{\boldsymbol{V}}{\delta}\right)}+\boldsymbol{\mu}-\sqrt{\frac{\boldsymbol{V}}{\delta}}\\ &= \boldsymbol{\mu}+\sqrt{\delta\boldsymbol{V}},\end{aligned}$$

where the first step holds since $\widetilde{\eta}^* = \boldsymbol{\mu} - \sqrt{\frac{\boldsymbol{V}}{\delta}} < 0$, and the fifth step is due to the definitions of $\boldsymbol{\mu}$ and $\boldsymbol{V}$.

Step (iii): Mean-dispersion form of the robust risk.

Therefore, by setting $\widetilde{\eta}^* = \boldsymbol{\mu} - \sqrt{\frac{\boldsymbol{V}}{\delta}}$, the dual objective (16) in its infimum form achieves the equality in (15), which is the upper bound of the primal problem (14) in its supremum form. Consequently, we obtain that

$$\mathcal{R}_{\mathbf{x}}^{-}(\boldsymbol{\theta}; P^{-}, \delta) = \mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}+\sqrt{\left\{\mathbb{E}_{P^{-}}\left[\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})-\mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\right\}\right]^2\right\}\cdot\delta}.$$

The proof is completed.

Case 2: KL-divergence. If the discrepancy metric d in (3) is chosen as the KL-divergence, then the robust risk $\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta}; P^{-}, \delta)$ is expressed as

$$\begin{aligned}\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta}; P^{-}, \delta) &= \sup_{Q^{-} \ll P^{-}} \left[\mathbb{E}_{Q^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\right\} : D_{\mathrm{KL}}(Q^{-} \| P^{-}) \leq \delta\right] \\ &= \sup_{Q^{-} \ll P^{-}} \left[\mathbb{E}_{Q^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\right\} : \mathbb{E}_{Q^{-}}\left\{\log\left(\frac{dQ^{-}}{dP^{-}}\right)\right\} \leq \delta\right]. \end{aligned} \tag{17}$$

By a change of measure and denoting the likelihood ratio $\mathsf{L}(\omega) \triangleq \frac{dQ^{-}(\omega)}{dP^{-}(\omega)}$ for $\omega \in \mathcal{X}$, the objective and the constraint in (17) can be expressed as

$$\begin{aligned}&\mathbb{E}_{Q^{-}}\Big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\Big\} = \mathbb{E}_{P^{-}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\frac{dQ^{-}}{dP^{-}}\right\} = \mathbb{E}_{P^{-}}\Big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\mathsf{L}(\mathbf{X}^{-})\Big\}; \\ &\mathbb{E}_{Q^{-}}\left\{\log\left(\frac{dQ^{-}}{dP^{-}}\right)\right\} = \mathbb{E}_{P^{-}}\left[\left\{\log\left(\frac{dQ^{-}}{dP^{-}}\right)\right\}\frac{dQ^{-}}{dP^{-}}\right] = \mathbb{E}_{P^{-}}\Big[\mathsf{L}(\mathbf{X}^{-})\log\left\{\mathsf{L}(\mathbf{X}^{-})\right\}\Big].\end{aligned}$$

Therefore, the expression of the robust risk $\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta}; P^{-}, \delta)$ can be rewritten as:

$$\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta}; P^{-}, \delta) = \begin{cases}\max\limits_{L \in \mathcal{L}} \ \mathbb{E}_{P^{-}}\Big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\mathsf{L}(\mathbf{X}^{-})\Big\} \\ \ s.t. \ \ \mathbb{E}_{P^{-}}\Big[\mathsf{L}(\mathbf{X}^{-})\log\left\{\mathsf{L}(\mathbf{X}^{-})\right\}\Big] \leq \delta,\end{cases} \tag{18}$$

where $\mathcal{L} = \{\mathsf{L} \in L^1(P^{-}) : \mathbb{E}_{P^{-}}\{\mathsf{L}(\mathbf{X}^{-})\} = 1, \mathsf{L} \geq 0 \ a.s.\}$. Since (18) is a convex optimization problem with respect to L, by introducing the Lagrange multiplier $\gamma > 0$, it can be further expressed as:

$$\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta}; P^{-}, \delta) = \max_{\mathsf{L} \in \mathcal{L}, \gamma \geq 0} \ \mathbb{E}_{P^{-}}\Big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\mathsf{L}(\mathbf{X}^{-})\Big\} - \gamma\Big\{\mathbb{E}_{P^{-}}\big[\mathsf{L}(\mathbf{X}^{-})\log\left\{\mathsf{L}(\mathbf{X}^{-})\right\}\big] - \delta\Big\}. \tag{19}$$

Step (i): Optimal form of the likelihood ratio L^*.

Suppose we can find $\gamma^* \geq 0$ and $\mathsf{L}^* \in \mathcal{L}$ such that L^* maximizes (19) for a fixed $\gamma = \gamma^*$ and $\mathbb{E}_{P^{-}}\Big[\mathsf{L}(\mathbf{X}^{-})\log\left\{\mathsf{L}(\mathbf{X}^{-})\right\}\Big] = \delta$. Then, for any L satisfying the constraint in (18), we have that

$$\begin{aligned}&\mathbb{E}_{P^{-}}\Big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\mathsf{L}^*(\mathbf{X}^{-})\Big\} \\ =&\mathbb{E}_{P^{-}}\Big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\mathsf{L}^*(\mathbf{X}^{-})\Big\} - \gamma^*\Big\{\mathbb{E}_{P^{-}}\big[\mathsf{L}^*(\mathbf{X}^{-})\log\left\{\mathsf{L}^*(\mathbf{X}^{-})\right\}\big] - \delta\Big\} \\ \geq&\mathbb{E}_{P^{-}}\Big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\mathsf{L}(\mathbf{X}^{-})\Big\} - \gamma^*\Big\{\mathbb{E}_{P^{-}}\big[\mathsf{L}(\mathbf{X}^{-})\log\left\{\mathsf{L}(\mathbf{X}^{-})\right\}\big] - \delta\Big\} \\ \geq&\mathbb{E}_{P^{-}}\Big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\mathsf{L}(\mathbf{X}^{-})\Big\},\end{aligned}$$

and hence, L^* is the optimal solution of (18).

We first assume the existence of such $\gamma^* \geq 0$ and consider the form of the corresponding L^*. Let $g(\mathsf{L}; \gamma) \triangleq \mathbb{E}_{P^{-}}\Big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\mathsf{L}(\mathbf{X}^{-})\Big\} - \gamma\Big\{\mathbb{E}_{P^{-}}\big[\mathsf{L}(\mathbf{X}^{-})\log\left\{\mathsf{L}(\mathbf{X}^{-})\right\}\big] - \delta\Big\}$ denote the objective function in (19). For a fixed $\gamma^* \in \mathbb{R}$, we consider the form of $\mathsf{L}^* \in \operatorname{argmax}_{\mathsf{L} \in \mathcal{L}} g(\mathsf{L}; \gamma^*)$, which can be expressed as

$$\begin{aligned}&\mathsf{L}^* \in \operatorname*{argmax}_{\mathsf{L} \in \mathcal{L}} \ \mathbb{E}_{P^{-}}\Big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\mathsf{L}(\mathbf{X}^{-})\Big\} - \gamma^*\Big\{\mathbb{E}_{P^{-}}\big[\mathsf{L}(\mathbf{X}^{-})\log\left\{\mathsf{L}(\mathbf{X}^{-})\right\}\big] - \delta\Big\} \\ \Leftrightarrow\ &\mathsf{L}^* \in \operatorname*{argmax}_{\mathsf{L} \in \mathcal{L}} \ -\gamma^*\Big(\mathbb{E}_{P^{-}}\Big\{-\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\mathsf{L}(\mathbf{X}^{-})/\gamma^*\Big\} + \mathbb{E}_{P^{-}}\Big[\mathsf{L}(\mathbf{X}^{-})\log\left\{\mathsf{L}(\mathbf{X}^{-})\right\}\Big]\Big) \\ \Leftrightarrow\ &\mathsf{L}^* dP^{-} \in \operatorname*{argmin}_{Q^{-} \in \mathcal{P}_p(\mathcal{X})} \ \mathbb{E}_{Q^{-}}\Big\{-\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})/\gamma^*\Big\} + D_{\mathrm{KL}}(Q^{-} \| P^{-})\Big].\end{aligned}$$

By Lemma 3, we obtain that

$$\mathsf{L}^*(\mathbf{X}^{-}) = \exp\left\{\frac{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})}{\gamma^*}\right\} \Big/ \mathbb{E}_{P^{-}}\left[\exp\left\{\frac{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})}{\gamma^*}\right\}\right]. \tag{20}$$

is the unique optimal solution of $\mathsf{L}^* \in \operatorname{argmax}_{\mathsf{L}\in\mathcal{L}} g(\mathsf{L};\gamma^*)$ for a fixed γ^* since the similarity measure $\mathcal{S}_{\boldsymbol{\theta}}$ is a bounded function.

Step (ii): Existence of γ^*.

If the γ^* in Step (i) exists, then the optimal L^* is given in (20), and the constraint and objective in (18) can be expressed as below:

$$\begin{aligned}
\delta &= \mathbb{E}_{P^-}\Big[\mathsf{L}^*(\mathbf{X}^-)\log\{\mathsf{L}^*(\mathbf{X}^-)\}\Big] \\
&= \mathbb{E}_{P^-}\left(\frac{\exp\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})/\gamma^*\}}{\mathbb{E}_{P^-}[\exp\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})/\gamma^*\}]}\cdot\left\{\frac{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})}{\gamma^*}-\log\mathbb{E}_{P^-}\left[\exp\left\{\frac{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})}{\gamma^*}\right\}\right]\right\}\right) \\
&= \frac{1}{\gamma^*}\cdot\frac{\mathbb{E}_{P^-}[\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})\cdot\exp\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})/\gamma^*\}]}{\mathbb{E}_{P^-}[\exp\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})/\gamma^*\}]}-\log\mathbb{E}_{P^-}\left[\exp\left\{\frac{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})}{\gamma^*}\right\}\right] \\
&= \bar{\varrho}\cdot\frac{\mathbb{E}_{P^-}[\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})\cdot\exp\{\bar{\varrho}\cdot\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})\}]}{\mathbb{E}_{P^-}[\exp\{\bar{\varrho}\cdot\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})\}]}-\log\mathbb{E}_{P^-}\Big[\exp\{\bar{\varrho}\cdot\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})\}\Big] \\
&\triangleq \bar{\varrho}\hbar'(\bar{\varrho})-\hbar(\bar{\varrho}); \qquad (21)
\end{aligned}$$

$$\mathbb{E}_{P^-}\Big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})\mathsf{L}^*(\mathbf{X}^-)\Big\} = \frac{\mathbb{E}_{P^-}[\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})\cdot\exp\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})/\gamma^*\}]}{\mathbb{E}_{P^-}[\exp\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})/\gamma^*\}]} = \hbar'(\bar{\varrho}), \qquad (22)$$

where we let $\varrho \triangleq 1/\gamma$, $\bar{\varrho} \triangleq 1/\gamma^*$, and $\hbar(\varrho) = \log\mathbb{E}_{P^-}[\exp\{\varrho\cdot\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})\}]$. Here $\hbar$ is the cumulant generating function of $\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})$, which is infinitely differentiable and strictly convex for non-constant $\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})$, and passes through the origin (Shalizi, 2006). Moreover, using a power series expansion, it can be expressed as: $\hbar(\varrho) = \sum_{j=1}^{\infty}\hbar^{(j)}(0)\,\varrho^j$, where $\hbar^{(j)}$ denotes the jth derivative of $\hbar$, and $\hbar^{(j)}(0)$ is referred to as the jth cumulant. It can be verified that $\hbar^{(1)}(0) = \mathbb{E}_{P^-}\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})\}$, $\hbar^{(2)}(0) = \mathbb{E}_{P^-}\Big\{\big[\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x}) - \mathbb{E}_{P^-}\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})\}\big]^2\Big\} > 0$, and $\hbar^{(3)}(0) = \mathbb{E}_{P^-}\Big\{\big[\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x}) - \mathbb{E}_{P^-}\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})\}\big]^3\Big\}$.

By the strict convexity of $\hbar$, we have that $d\{\varrho\hbar'(\varrho)-\hbar(\varrho)\}/d\varrho = \hbar''(\varrho) > 0$, and hence $\varrho\hbar'(\varrho) - \hbar(\varrho)$ is strictly increasing in ϱ. Moreover, by (21), using Taylor's expansion, we obtain that

$$\begin{aligned}
\delta &= \bar{\varrho}\,\hbar'(\bar{\varrho}) - \hbar(\bar{\varrho}) \\
&= \bar{\varrho}\sum_{j=0}^{+\infty}\frac{1}{j!}\hbar^{(j+1)}(0)\,\bar{\varrho}^j - \sum_{j=0}^{+\infty}\frac{1}{j!}\hbar^{(j)}(0)\,\bar{\varrho}^j \\
&= \sum_{j=1}^{+\infty}\frac{1}{(j-1)!}\hbar^{(j)}(0)\,\bar{\varrho}^j - \sum_{j=1}^{+\infty}\frac{1}{j!}\hbar^{(j)}(0)\,\bar{\varrho}^j \\
&= \sum_{j=1}^{+\infty}\left\{\frac{1}{(j-1)!}-\frac{1}{j!}\right\}\hbar^{(j)}(0)\,\bar{\varrho}^j \\
&= \frac{1}{2}\hbar^{(2)}(0)\,\bar{\varrho}^2 + \frac{1}{3}\hbar^{(3)}(0)\,\bar{\varrho}^3 + \mathcal{O}(\bar{\varrho}^4). \qquad (23)
\end{aligned}$$

Since $\hbar^{(2)}(0) > 0$ and the remainder is continuous in ϱ, we have that there exists a small $\bar{\varrho}$ satisfying the equation (23) for a small enough δ, and that $\bar{\varrho}$ is the unique solution of (21). Correspondingly, for $\gamma^* = 1/\bar{\varrho}$, the associated L^* satisfies the constraint $\mathbb{E}_{P^-}\Big[\mathsf{L}^*(\mathbf{X}^-)\log\{\mathsf{L}^*(\mathbf{X}^-)\}\Big] = \delta$. Hence, $\mathcal{R}^-_{\mathbf{x}}(\boldsymbol{\theta};P^-,\delta) = \mathbb{E}_{P^-}\Big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-;\mathbf{x})\mathsf{L}^*(\mathbf{X}^-)\Big\}$.

Step (iii): Mean-dispersion form of the robust risk.

Now, we examine the form of the robust risk. By (23), we have

$$\frac{2\delta}{\hbar^{(2)}(0)} = \bar{\varrho}^2 + \frac{2\hbar^{(3)}(0)}{3\hbar^{(2)}(0)}\bar{\varrho}^3 + \mathcal{O}(\bar{\varrho}^4) = \bar{\varrho}^2\left\{1+\frac{2\hbar^{(3)}(0)}{3\hbar^{(2)}(0)}\bar{\varrho}+\mathcal{O}(\bar{\varrho}^2)\right\},$$

and further obtain that

$$
\begin{aligned}
\bar{\varrho} =& \sqrt{\frac{2\delta}{\hbar^{(2)}(0)}} \cdot \sqrt{1 \Big/ \left\{1 + \frac{2\hbar^{(3)}(0)}{3\hbar^{(2)}(0)}\bar{\varrho} + \mathcal{O}(\bar{\varrho}^2)\right\}} \\
=& \sqrt{\frac{2\delta}{\hbar^{(2)}(0)}} \cdot \sqrt{1 - \frac{2\hbar^{(3)}(0)}{3\hbar^{(2)}(0)}\bar{\varrho} + \mathcal{O}(\bar{\varrho}^2)} \\
=& \sqrt{\frac{2\delta}{\hbar^{(2)}(0)}} \cdot \left\{1 - \frac{\hbar^{(3)}(0)}{3\hbar^{(2)}(0)}\bar{\varrho} + \mathcal{O}(\bar{\varrho}^2)\right\} \\
=& \sqrt{\frac{2\delta}{\hbar^{(2)}(0)}} - \frac{2\hbar^{(3)}(0)}{3\{\hbar^{(2)}(0)\}^2}\delta + \mathcal{O}(\delta).
\end{aligned}
$$

Hence, by (22), we have that

$$
\begin{aligned}
\mathcal{R}_{\mathbf{x}}^{-}(\boldsymbol{\theta}; P^{-}, \delta) =& \mathbb{E}_{P^{-}}\Big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\mathsf{L}^*(\mathbf{X}^{-})\Big\} \\
=& \hbar'(\bar{\varrho}) = \hbar^{(1)}(0) + \hbar^{(2)}(0)\bar{\varrho} + \frac{\hbar^{(3)}(0)}{2}\bar{\varrho}^2 + \mathcal{O}(\bar{\varrho}^2) \\
=& \hbar^{(1)}(0) + \sqrt{2\hbar^{(2)}(0)\delta} + \mathcal{O}(\delta) \\
=& \mathbb{E}_{P^{-}}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\big\} + \sqrt{2\mathbb{E}_{P^{-}}\left\{\big[\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x}) - \mathbb{E}_{P^{-}}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\big\}\big]^2\right\}\delta} + \mathcal{O}(\delta).
\end{aligned}
$$

Therefore, the proof is established.

Case 3: p-Wasserstein distance. If the discrepancy metric $\mathscr{d}$ in (3) is chosen as the p-Wasserstein distance, then the robust risk $\mathcal{R}_{\mathbf{x}}^{-}(\boldsymbol{\theta}; P^{-}, \delta)$ is expressed as

$$
\mathcal{R}_{\mathbf{x}}^{-}(\boldsymbol{\theta}; P^{-}, \delta) = \sup_{Q^{-} \in \mathcal{P}(\Omega)} \left[\mathbb{E}_{Q^{-}}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\big\} : W_p(Q^{-}, P^{-}) \leq \delta\right]. \tag{24}
$$

Let $\Delta\mathcal{R}_{\mathbf{x}}^{-} \triangleq \mathcal{R}_{\mathbf{x}}^{-}(\boldsymbol{\theta}; P^{-}, \delta) - \mathbb{E}_{P^{-}}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\big\}$ denote the difference of the robust risk and the nominal risk. By Lemma 1, we have that

$$
\begin{aligned}
\Delta\mathcal{R}_{\mathbf{x}}^{-} =& \min_{\gamma \geq 0}\left\{\gamma\delta^p + \mathbb{E}_{P^{-}}\left[\sup_{\widetilde{\mathbf{x}}^{-} \in \Omega}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\widetilde{\mathbf{x}}^{-}; \mathbf{x}) - \gamma\|\widetilde{\mathbf{x}}^{-} - \mathbf{X}^{-}\|^p\big\}\right]\right\} - \mathbb{E}_{P^{-}}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\big\} \\
=& \min_{\gamma \geq 0}\left(\gamma\delta^p + \mathbb{E}_{P^{-}}\left\{\sup_{\widetilde{\mathbf{x}}^{-} \in \Omega}\left[\Big\{\mathcal{S}_{\boldsymbol{\theta}}(\widetilde{\mathbf{x}}^{-}; \mathbf{x}) - \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\Big\} - \gamma\|\widetilde{\mathbf{x}}^{-} - \mathbf{X}^{-}\|^p\right]\right\}\right).
\end{aligned} \tag{25}
$$

Step (i): Upper bound on $\mathcal{S}_{\boldsymbol{\theta}}(\widetilde{\mathbf{x}}^{-}; \mathbf{x}) - \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^{-}; \mathbf{x})$.

For any $\widetilde{\mathbf{x}}^{-}, \mathbf{x}^{-} \in \mathcal{X}$, by the mean value theorem, there exists $\check{\mathbf{x}}^{-} \in \mathcal{X}$ between $\widetilde{\mathbf{x}}^{-}$ and $\mathbf{x}^{-}$ such that

$$
\mathcal{S}_{\boldsymbol{\theta}}(\widetilde{\mathbf{x}}^{-}; \mathbf{x}) - \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^{-}; \mathbf{x}) = \langle \nabla\mathcal{S}_{\boldsymbol{\theta}}(\check{\mathbf{x}}^{-}; \mathbf{x}), \widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\rangle,
$$

which implies that

$$
\begin{aligned}
& |\mathcal{S}_{\boldsymbol{\theta}}(\widetilde{\mathbf{x}}^{-}; \mathbf{x}) - \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^{-}; \mathbf{x}) - \langle \nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^{-}; \mathbf{x}), \widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\rangle| \\
=& |\langle \nabla\mathcal{S}_{\boldsymbol{\theta}}(\check{\mathbf{x}}^{-}; \mathbf{x}) - \nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^{-}; \mathbf{x}), \widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\rangle| \\
\leq& \|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\check{\mathbf{x}}^{-}; \mathbf{x}) - \nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^{-}; \mathbf{x})\|_* \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\| \\
\leq& \|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\widetilde{\mathbf{x}}^{-}; \mathbf{x}) - \nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^{-}; \mathbf{x})\|_* \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|,
\end{aligned} \tag{26}
$$

where the inequality in the penultimate step is due to the Cauchy–Schwarz inequality.

If $\|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\| \leq \eta_0$, by the smoothness condition (b), we have that

$$
\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\widetilde{\mathbf{x}}^{-}; \mathbf{x}) - \nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^{-}; \mathbf{x})\|_* \leq \mathscr{M}_3\|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|. \tag{27}
$$

If $\|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\| \geq \eta_0$, by the smoothness condition (a), we have that

$$\|\nabla \mathcal{S}_{\boldsymbol{\theta}}(\widetilde{\mathbf{x}}^{-};\mathbf{x}) - \nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^{-};\mathbf{x})\|_* \leq \mathscr{M}_1 + \mathscr{M}_2 \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|^{\zeta - 1}. \tag{28}$$

Combining (26), (27) and (28), we further obtain that

$$\begin{aligned}
&|\mathcal{S}_{\boldsymbol{\theta}}(\widetilde{\mathbf{x}}^{-};\mathbf{x}) - \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^{-};\mathbf{x}) - < \nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^{-};\mathbf{x}), \widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-} > | \\
&= \mathbf{1}(\|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\| \leq \eta_0) \cdot \mathscr{M}_3 \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|^2 + \mathbf{1}(\|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\| \geq \eta_0) \cdot \left(\mathscr{M}_1 \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\| + \mathscr{M}_2 \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|^{\zeta}\right) \\
&\triangleq \mathcal{I}_1 + \mathcal{I}_2,
\end{aligned}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function, $\mathcal{I}_1 \triangleq \mathbf{1}(\|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\| \leq \eta_0) \cdot \mathscr{M}_3 \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|^2$ and $\mathcal{I}_2 \triangleq \mathbf{1}(\|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\| \geq \eta_0) \cdot \left(\mathscr{M}_1 \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\| + \mathscr{M}_2 \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|^{\zeta}\right)$.

For $\mathcal{I}_1$, if $1 \leq p \leq 2$, we have

$$\begin{aligned}
\mathcal{I}_1 &\leq \mathbf{1}(\|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\| \leq \eta_0) \cdot \mathscr{M}_3 \left(\frac{\eta_0}{\|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|}\right)^{2-p} \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|^2 \\
&\leq \mathscr{M}_3 \eta_0^{2-p} \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|^p.
\end{aligned}$$

If $p > 2$, we have $\mathcal{I}_1 \leq \mathscr{M}_3 \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|^2$. For $\mathcal{I}_2$, we have the following upper bound:

$$\begin{aligned}
\mathcal{I}_2 \leq &\mathbf{1}(\|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\| \geq \eta_0) \cdot \left\{\mathscr{M}_1 \left(\frac{\|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|}{\eta_0}\right)^{p-1} \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\| + \mathscr{M}_2 \left(\frac{\|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|}{\eta_0}\right)^{p-\zeta} \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|^{\zeta}\right\} \\
\leq &\left(\mathscr{M}_1 \eta_0^{-(p-1)} + \mathscr{M}_2 \eta_0^{-(p-\zeta)}\right) \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|^p.
\end{aligned}$$

Combining the discussion above, we have that

$$\begin{aligned}
&|\mathcal{S}_{\boldsymbol{\theta}}(\widetilde{\mathbf{x}}^{-};\mathbf{x}) - \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^{-};\mathbf{x}) - < \nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^{-};\mathbf{x}), \widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-} > | \\
&\leq \begin{cases} \bar{\mathscr{M}}\ \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|^p, \text{ if } 1 \leq p \leq 2; \\ \bar{\mathscr{M}}\ \left(\|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|^p + \|\widetilde{\mathbf{x}}^{-} - \mathbf{x}^{-}\|^2\right), \text{ if } p > 2, \end{cases}
\end{aligned} \tag{29}$$

where $\bar{\mathscr{M}} \triangleq \max\{\mathscr{M}_3 \eta_0^{2-p}, \mathscr{M}_3, \left(\mathscr{M}_1 \eta_0^{-(p-1)} + \mathscr{M}_2 \eta_0^{-(p-\zeta)}\right)\}$.

Step (ii): Mean-dispersion form of the robust risk when $p \in [1, 2]$.

When $p \in [1, 2]$, by (25) and (29), we have that

$$\begin{aligned}
\Delta \mathcal{R}_{\mathbf{x}}^{-} &\leq \min_{\gamma \geq 0} \left(\gamma \delta^p + \mathbb{E}_{P^{-}} \left\{\sup_{\widetilde{\mathbf{x}}^{-} \in \Omega} \left[\left\{\langle \nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}), \widetilde{\mathbf{x}}^{-} - \mathbf{X}^{-}\rangle + \bar{\mathscr{M}}\ \|\widetilde{\mathbf{x}}^{-} - \mathbf{X}^{-}\|^p\right\} - \gamma \|\widetilde{\mathbf{x}}^{-} - \mathbf{X}^{-}\|^p\right]\right\}\right) \\
&= \min_{\gamma \geq 0} \left\{\gamma \delta^p + \mathbb{E}_{P^{-}} \left[\sup_{\widetilde{\mathbf{x}}^{-} \in \Omega} \left\{\langle \nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}), \widetilde{\mathbf{x}}^{-} - \mathbf{X}^{-}\rangle - (\gamma - \bar{\mathscr{M}}) \|\widetilde{\mathbf{x}}^{-} - \mathbf{X}^{-}\|^p\right\}\right]\right\} \\
&\leq \min_{\gamma \geq 0} \left\{\gamma \delta^p + \mathbb{E}_{P^{-}} \left[\sup_{\widetilde{\mathbf{x}}^{-} \in \Omega} \left\{\|\nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_* \|\widetilde{\mathbf{x}}^{-} - \mathbf{X}^{-}\| - (\gamma - \bar{\mathscr{M}}) \|\widetilde{\mathbf{x}}^{-} - \mathbf{X}^{-}\|^p\right\}\right]\right\} \\
&= \min_{\gamma \geq -\bar{\mathscr{M}}} \left\{\gamma \delta^p + \mathbb{E}_{P^{-}} \left[\sup_{t \geq 0} \left\{\|\nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_* t - \gamma t^p\right\}\right]\right\} + \bar{\mathscr{M}} \delta^p \\
&\leq \min_{\gamma \geq 0} \left\{\gamma \delta^p + \mathbb{E}_{P^{-}} \left[\sup_{t \geq 0} \left\{\|\nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_* t - \gamma t^p\right\}\right]\right\} + \bar{\mathscr{M}} \delta^p \\
&\triangleq \mathcal{I}_4 + \bar{\mathscr{M}} \delta^p,
\end{aligned} \tag{30}$$

where $\mathcal{I}_4 \triangleq \min_{\gamma \geq 0} \left\{\gamma \delta^p + \mathbb{E}_{P^{-}} \left[\sup_{t \geq 0} \left\{\|\nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_* t - \gamma t^p\right\}\right]\right\}$ in (30) and (**??**), and the third step is due to the Cauchy–Schwarz inequality.

By taking the derivative with respect to t in the supremum in $\mathcal{I}_4$ and setting it to zero, we obtain that the optimal value of t is $t^* = \{\|\nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_* / (\gamma p)\}^{1/(p-1)}$. Let q denote the Hölder number of p, that is $\frac{1}{p}+\frac{1}{q}=1$. Then, $q=\frac{p}{p-1}$ and $\frac{q}{p}=\frac{1}{p-1}$. We have that

$$\begin{aligned}
&\sup_{t\geq 0}\left\{\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}t-\gamma t^{p}\right\}\\
=&\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}t^{*}-\gamma(t^{*})^{p}\\
=&\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}\cdot\left\{\frac{\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}}{\gamma p}\right\}^{\frac{1}{p-1}}-\gamma\cdot\left\{\frac{\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}}{\gamma p}\right\}^{\frac{p}{p-1}}\\
=&\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}^{\frac{p}{p-1}}(\gamma p)^{-\frac{1}{p-1}}-\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}^{\frac{p}{p-1}}\gamma^{-\frac{1}{p-1}}p^{-\frac{p}{p-1}}\\
=&\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}^{q}(\gamma p)^{-\frac{1}{p-1}}\left(1-\frac{1}{p}\right).
\end{aligned}$$

Thus, we further obtain that

$$\mathcal{I}_{4}=\min_{\gamma\geq 0}\left[\gamma\delta^{p}+\left(1-\frac{1}{p}\right)(\gamma p)^{-\frac{1}{p-1}}\mathbb{E}_{P^{-}}\left\{\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}^{q}\right\}\right].$$

Similarly, by taking the derivative with respect to γ in the infimum and set it to zero, we obtain that the optimal value of γ is $\gamma^{*}=\frac{1}{p}\delta^{-(p-1)}\left\{\mathbb{E}_{P^{-}}\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}^{q}\right\}^{1/q}$. Hence, by substituting γ^{*} into the previous expression and simplifying the formula, we further obtain that

$$\begin{aligned}
\mathcal{I}_{4}=&\frac{1}{p}\delta^{-(p-1)}\left\{\mathbb{E}_{P^{-}}\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}^{q}\right\}^{1/q}\delta^{p}\\
&+\left\{\frac{1}{p}\delta^{-(p-1)}\left\{\mathbb{E}_{P^{-}}\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}^{q}\right\}^{1/q}\right\}^{-\frac{1}{p-1}}\left(\frac{p-1}{p}\right)p^{-\frac{1}{p-1}}\\
=&\frac{1}{p}\delta\left\{\mathbb{E}_{P^{-}}\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}^{q}\right\}^{1/q}+\left(\frac{p-1}{p}\right)\delta\left\{\mathbb{E}_{P^{-}}\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}^{q}\right\}^{1/q}\\
=&\delta\left\{\mathbb{E}_{P^{-}}\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}^{q}\right\}^{1/q}.
\end{aligned}\tag{31}$$

Combining (30) and (31), we obtain that

$$\Delta\mathcal{R}_{\mathbf{x}}^{-}\leq\delta\left\{\mathbb{E}_{P^{-}}\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}^{q}\right\}^{1/q}+\bar{\mathcal{M}}\delta^{p}.$$

Step (iii): Mean-dispersion form of the robust risk when $p\in(2,\infty)$.

When $p\in(2,\infty)$, by (25) and (29), similar to (30) in Step (ii), we have that

$$\begin{aligned}
\Delta\mathcal{R}_{\mathbf{x}}^{-}\leq&\min_{\gamma\geq 0}\left(\gamma\delta^{p}+\mathbb{E}_{P^{-}}\left\{\sup_{\widetilde{\mathbf{x}}^{-}\in\Omega}\left[\left\{\langle\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}),\widetilde{\mathbf{x}}^{-}-\mathbf{X}^{-}\rangle\right.\right.\right.\right.\\
&\left.\left.\left.\left.+\bar{\mathcal{M}}(\|\widetilde{\mathbf{x}}^{-}-\mathbf{X}^{-}\|^{p}+\|\widetilde{\mathbf{x}}^{-}-\mathbf{X}^{-}\|^{2})\right\}-\gamma\|\widetilde{\mathbf{x}}^{-}-\mathbf{X}^{-}\|^{p}\right]\right\}\right)\\
\leq&\min_{\gamma\geq 0}\left\{\gamma\delta^{p}+\mathbb{E}_{P^{-}}\left[\sup_{\widetilde{\mathbf{x}}^{-}\in\Omega}\left\{\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}\|\widetilde{\mathbf{x}}^{-}-\mathbf{X}^{-}\|\right.\right.\right.\\
&\left.\left.\left.+\bar{\mathcal{M}}\|\widetilde{\mathbf{x}}^{-}-\mathbf{X}^{-}\|^{p}+\bar{\mathcal{M}}\|\widetilde{\mathbf{x}}^{-}-\mathbf{X}^{-}\|^{2}-\gamma\|\widetilde{\mathbf{x}}^{-}-\mathbf{X}^{-}\|^{p}\right\}\right]\right\}\\
=&\min_{\gamma\geq 0}\left\{\gamma\delta^{p}+\mathbb{E}_{P^{-}}\left[\sup_{t\geq 0}\left\{\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}t+\bar{\mathcal{M}}t^{p}+\bar{\mathcal{M}}t^{2}-\gamma t^{p}\right\}\right]\right\}\\
\leq&\min_{\gamma\geq 0}\left\{\gamma\delta^{p}+\mathbb{E}_{P^{-}}\left[\sup_{t\geq 0}\left\{\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}t+\bar{\mathcal{M}}t^{2}-\gamma t^{p}\right\}\right]\right\}+\bar{\mathcal{M}}\delta^{p}\\
=&\min_{\gamma_{1},\gamma_{2}\geq 0}\left\{(\gamma_{1}+\gamma_{2})\delta^{p}+\mathbb{E}_{P^{-}}\left[\sup_{t\geq 0}\left\{\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}t+\bar{\mathcal{M}}t^{2}-(\gamma_{1}+\gamma_{2})t^{p}\right\}\right]\right\}+\bar{\mathcal{M}}\delta^{p}\\
\leq&\min_{\gamma_{1}\geq 0}\left\{\gamma_{1}\delta^{p}+\mathbb{E}_{P^{-}}\left[\sup_{t\geq 0}\left\{\|\nabla\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\|_{*}t-\gamma_{1}t^{p}\right\}\right]\right\}+\min_{\gamma_{2}\geq 0}\left\{\gamma_{2}\delta^{p}+\sup_{t\geq 0}\left(\bar{\mathcal{M}}t^{2}-\gamma_{2}t^{p}\right)\right\}+\bar{\mathcal{M}}\delta^{p}\\
\triangleq&\mathcal{I}_{5}+\mathcal{I}_{6}+\bar{\mathcal{M}}\delta^{p}
\end{aligned}\tag{32}$$

where $\mathcal{I}_5 \triangleq \min_{\gamma_1 \geq 0} \left\{ \gamma_1 \delta^p + \mathbb{E}_{P^-} \left[\sup_{t \geq 0} \left\{ \|\nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x})\|_* t - \gamma_1 t^p \right\} \right] \right\}$, and $\mathcal{I}_6 \triangleq \min_{\gamma_2 \geq 0} \left\{ \gamma_2 \delta^p + \sup_{t \geq 0} \left(\bar{\mathscr{M}} t^2 - \gamma_2 t^p \right) \right\}$.

For $\mathcal{I}_5$, similar to the discussion on $\mathcal{I}_4$ with $p \in [1, 2]$ as in (31), we obtain that, for $p \in (2, \infty)$,

$$\mathcal{I}_5 = \delta \left\{ \mathbb{E}_{P^-} \|\nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x})\|_*^q \right\}^{1/q}. \tag{33}$$

For $\mathcal{I}_6$, by taking the derivative with respect to t in the supremum and setting it to zero, we obtain that the optimal value of t is given by $t^* = \left\{ 2\bar{\mathscr{M}}/(\gamma_2 p) \right\}^{1/(p-2)}$. Then,

$$\begin{aligned}
\mathcal{I}_6 &= \min_{\gamma_2 \geq 0} \left\{ \gamma_2 \delta^p + \bar{\mathscr{M}} (t^*)^2 - \gamma_2 (t^*)^p \right\} \\
&= \min_{\gamma_2 \geq 0} \left\{ \gamma_2 \delta^p + \bar{\mathscr{M}} \cdot \left(\frac{2\bar{\mathscr{M}}}{\gamma_2 p} \right)^{\frac{2}{p-2}} - \gamma_2 \left(\frac{2\bar{\mathscr{M}}}{\gamma_2 p} \right)^{\frac{p}{p-2}} \right\} \\
&= \min_{\gamma_2 \geq 0} \left\{ \gamma_2 \delta^p + \left(\frac{\gamma_2 p}{2} \right)^{-\frac{2}{p-2}} \bar{\mathscr{M}}^{\frac{p}{p-2}} - \gamma_2^{-\frac{2}{p-2}} \cdot \left(\frac{p}{2} \right)^{-\frac{2}{p-2}} \cdot \left(\frac{p}{2} \right)^{-1} \cdot \bar{\mathscr{M}}^{\frac{p}{p-2}} \right\} \\
&= \min_{\gamma_2 \geq 0} \left\{ \gamma_2 \delta^p + \frac{p-2}{p} \left(\frac{\gamma_2 p}{2} \right)^{-\frac{2}{p-2}} \bar{\mathscr{M}}^{\frac{p}{p-2}} \right\}.
\end{aligned}$$

By taking the derivative with respect to γ_2, we further obtain that the optimal value of γ_2 is $\gamma_2^* = \bar{\mathscr{M}} \delta^{-(p-2)} \left(\frac{p}{2} \right)^{-1}$, and that

$$\mathcal{I}_6 = \gamma_2^* \delta^p + \frac{p-2}{p} \left(\frac{\gamma_2^* p}{2} \right)^{-\frac{2}{p-2}} \bar{\mathscr{M}}^{\frac{p}{p-2}} = \bar{\mathscr{M}} \delta^2. \tag{34}$$

Combining (33), (34), and (34), we obtain

$$\Delta \mathcal{R}_{\mathbf{x}}^- \leq \delta \left\{ \mathbb{E}_{P^-} \|\nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x})\|_*^q \right\}^{1/q} + \bar{\mathscr{M}} \delta^2 + \bar{\mathscr{M}} \delta^p. \tag{35}$$

Hence, the proof is completed.

□

Theorem A.1. *For the contaminated training distribution P_{train}^-, suppose that the induced distribution of $\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}^-)$ is non-degenerate. Let s^* represent the $1 - \epsilon$ quantile of this distribution, such that $P_{train}^- \left\{ \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) \leq s^* \right\} = 1 - \epsilon$. Let p_{train}^- denote the density / mass function of P_{train}^-. We define the following truncated distribution:*

$$p^*(\mathbf{x}^-) \triangleq \begin{cases} \dfrac{1}{1-\epsilon} p_{train}^-(\mathbf{x}^-), & \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^-; \mathbf{x}) \leq s^*; \\ 0, & \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^-; \mathbf{x}) > s^*. \end{cases}$$

Let P^* denote the associated probability measure of p^*. Let $\Re_1 \triangleq \frac{1}{1-\epsilon} \int_0^{s^*} s \, dP_{train}^- \left\{ \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) \leq s \right\}$ and $\Re_1 \triangleq \frac{1}{1-\epsilon} \int_0^{s^*} s^2 \, dP_{train}^- \left\{ \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) \leq s \right\}$. For different choices of the discrepancy metric $\mathscr{d}$ in (3), we have the following upper bounds on the outlier robust risk $\mathcal{R}_{\mathbf{x}}^-(\boldsymbol{\theta}; P_{train}^-, \delta, \epsilon)$ given in (4).

(i) If $\mathscr{d}$ is the χ^2-divergence, then for a small enough δ,

$$\mathcal{R}_{\mathbf{x}}^-(\boldsymbol{\theta}; P^-, \delta) \leq \mathbb{E}_{P^*} \left\{ \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) \right\} + \sqrt{\delta \mathbb{V}_{P^*} \left\{ \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) \right\}},$$

where $\mathbb{E}_{P^*} \left\{ \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) \right\} = \Re_1$, and $\mathbb{V}_{P^*} \left\{ \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) \right\} = \Re_2 - \Re_1^2$.

(ii) If $\mathscr{d}$ is the KL-divergence, then for a small enough δ,

$$\mathcal{R}_{\mathbf{x}}^-(\boldsymbol{\theta}; P^-, \delta) \leq \mathbb{E}_{P^*} \left\{ \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) \right\} + \sqrt{2\delta \mathbb{V}_{P^*} \left\{ \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) \right\}} + \mathcal{O}(\delta),$$

where $\mathbb{E}_{P^*} \left\{ \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) \right\} = \Re_1$, and $\mathbb{V}_{P^*} \left\{ \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) \right\} = \Re_2 - \Re_1^2$.

(iii) Suppose d is the p-Wasserstein distance with $p \in [1, +\infty)$ and the cost function $c(\cdot,\cdot)$ in Definition A.1 is chosen as a norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$. Assume the following smoothness condition are true.

a. For any $\widetilde{\mathbf{x}}^-, \mathbf{x}^-, \mathbf{x} \in \mathcal{X}$, $\exists \mathscr{M}_1, \mathscr{M}_2 > 0$ and $\zeta \in [1, p]$, such that $\|\nabla \mathcal{S}_{\boldsymbol{\theta}}(\widetilde{\mathbf{x}}^-; \mathbf{x}) - \nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^-; \mathbf{x})\|_* \le \mathscr{M}_1 + \mathscr{M}_2 \|\widetilde{\mathbf{x}}^- - \mathbf{x}^-\|^{\zeta - 1}$.

b. There exists $\eta_0 > 0$ and $\mathscr{M}_3 > 0$, such that for any $\widetilde{\mathbf{x}}^-, \mathbf{x}^-, \mathbf{x} \in \mathcal{X}$, if $\|\widetilde{\mathbf{x}}^- - \mathbf{x}^-\| \le \eta_0$, then $\|\nabla \mathcal{S}_{\boldsymbol{\theta}}(\widetilde{\mathbf{x}}^-; \mathbf{x}) - \nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^-; \mathbf{x})\|_* \le \mathscr{M}_3 \|\widetilde{\mathbf{x}}^- - \mathbf{x}^-\|$.

Let q denote the Hölder number of p, that is $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$\mathcal{R}_{\mathbf{x}}^-(\boldsymbol{\theta}; P^-, \delta) \le \mathbb{E}_{P^*}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x})\big\} + \delta \left\{\mathbb{E}_{P^*} \|\nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x})\|_*^q\right\}^{1/q} + \mathcal{O}(\delta^{2 \wedge p}),$$

where $\mathbb{E}_{P^*}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x})\big\} = \Re_1$.

Proof. We first examine form of the outlier robust risk given in (4) when different choices of the discrepancy metric d in (3). Proof techniques in Zhai et al. (2021) are used.

Case 1: χ^2-divergence. If the discrepancy metric d in (3) is chosen as the χ^2-divergence, by (4) and Lemma 5, we have that

$$\begin{aligned}
\mathcal{R}_{\mathbf{x}}^-(\boldsymbol{\theta}; P_{\text{train}}^-, \delta, \epsilon) &= \inf_{P' \in \mathcal{P}_p(\mathcal{X})} \Big\{\mathcal{R}_{\mathbf{x}}^-(\boldsymbol{\theta}; P', \delta) : \exists \widetilde{P}' \in \mathcal{P}_p(\mathcal{X}) \; s.t. \; P_{\text{train}}^- = (1-\epsilon)P' + \epsilon \widetilde{P}'\Big\} \\
&= \inf_{P' \in \mathcal{P}_p(\mathcal{X})} \Big\{\mathbb{E}_{P'}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x})\big\} + \sqrt{\delta \mathbb{V}_{P'}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x})\big\}} : \\
&\qquad \exists \widetilde{P}' \in \mathcal{P}_p(\mathcal{X}) \; s.t. \; P_{\text{train}}^- = (1-\epsilon)P' + \epsilon \widetilde{P}'\Big\}
\end{aligned} \tag{36}$$

We consider the following quantity:

$$\begin{aligned}
\Re_1 &\triangleq \inf_{P' \in \mathcal{P}_p(\mathcal{X})} \Big\{\mathbb{E}_{P'}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x})\big\} : \exists \widetilde{P}' \in \mathcal{P}_p(\mathcal{X}) \; s.t. \; P_{\text{train}}^- = (1-\epsilon)P' + \epsilon \widetilde{P}'\Big\} \\
&= \inf_{P' \in \mathcal{P}_p(\mathcal{X})} \Big\{ \int_0^{+\infty} \left[1 - P'\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) \le s\right\}\right] ds : \exists \widetilde{P}' \in \mathcal{P}_p(\mathcal{X}) \; s.t. \; P_{\text{train}}^- = (1-\epsilon)P' + \epsilon \widetilde{P}'\Big\},
\end{aligned} \tag{37}$$

where the second equality holds since for a nonnegative random variable Z with cumulative distribution function F, if its kth moment $\mathbb{E}_F(Z^k)$ exists, then, it can be expressed as $\mathbb{E}_F(Z^k) = k\int_0^{+\infty} u^{k-1}\{1 - F(u)\}du$.

Since $P_{\text{train}}^- = (1-\epsilon)P' + \epsilon \widetilde{P}'$, we have that for any $s \ge 0$,

$$P'\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) \le s\right\} \le \min\left\{\frac{1}{1-\epsilon} P_{\text{train}}^- \left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) \le s\right\}, 1\right\}. \tag{38}$$

As in Zhai et al. (2021), we show the equality in (38) can be achieved by some $P^* \in \mathcal{P}_p(\mathcal{X})$. Specifically, since P_{train}^- and $\mathcal{S}_{\boldsymbol{\theta}}$ are continuous, there exists an s^* such that $P_{\text{train}}^- \left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) > s^*\right\} = \epsilon$. Define

$$p^*(\mathbf{x}^-) \triangleq \begin{cases} \dfrac{1}{1-\epsilon} p_{\text{train}}^-(\mathbf{x}^-), & \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^-; \mathbf{x}) \le s^*; \\ 0, & \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^-; \mathbf{x}) > s^*, \end{cases} \tag{39}$$

where p_{train}^- represents the density / mass function of P_{train}^-. Let P^* denote the associated measure of p^*. For the P^* defined above, we have $\int_{\mathcal{X}} dP^*(\mathbf{x}^-) = \frac{1}{1-\epsilon} \int_{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}^-; \mathbf{x}) \le s^*} dP_{\text{train}}^-(\mathbf{x}^-) = \frac{1}{1-\epsilon} P_{\text{train}}^- \left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^-; \mathbf{x}) \le s^*\right\} = 1$. Therefore, P^* defined in (39) is probability distribution achieving the equality in (38).

Thus, by substituting P^* into (37) and utilizing (38), $\Re_1$ can be written as below:

$$\begin{aligned}
\Re_1 =& \mathbb{E}_{P^*}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\big\} \\
=& \int_0^{+\infty} \left[1 - P^*\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}) \le s\right\}\right] ds \\
=& \int_0^{+\infty} \left[1 - \frac{1}{1-\epsilon} P^{-}_{\text{train}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}) \le s\right\}\right] \mathbf{1}\left[P^{-}_{\text{train}}\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}) \le s\} \le 1-\epsilon\right] ds \\
=& \int_0^{+\infty} \left[1 - \frac{1}{1-\epsilon} P^{-}_{\text{train}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}) \le s\right\}\right] \mathbf{1}(s \le s^*) ds \\
=& \frac{1}{1-\epsilon}\left[(1-\epsilon)s^* - \int_0^{s^*} P^{-}_{\text{train}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}) \le s\right\} ds\right] \\
=& \frac{1}{1-\epsilon}\left\{\left[s\, P^{-}_{\text{train}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}) \le s\right\}\right]\Big|_0^{s^*} - \int_0^{s^*} P^{-}_{\text{train}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}) \le s\right\} ds\right\} \\
=& \frac{1}{1-\epsilon}\int_0^{s^*} s\, dP^{-}_{\text{train}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}) \le s\right\}
\end{aligned} \tag{40}$$

For the variance term in (36), we consider the following 2nd order moment:

$$\begin{aligned}
\Re_2 \triangleq& \mathbb{E}_{P^*}\left[\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\big\}^2\right] \\
=& 2\int_0^{+\infty} s\ \left[1 - P^*\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}) \le s\right\}\right] ds \\
=& \int_0^{+\infty} 2s \cdot \left[1 - \frac{1}{1-\epsilon} P^{-}_{\text{train}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}) \le s\right\}\right] \mathbf{1}(s \le s^*) ds \\
=& \frac{1}{1-\epsilon}\left[(1-\epsilon)(s^*)^2 - \int_0^{s^*} 2s P^{-}_{\text{train}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}) \le s\right\} ds\right] \\
=& \frac{1}{1-\epsilon}\left\{\left[s^2\, P^{-}_{\text{train}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}) \le s\right\}\right]\Big|_0^{s^*} - \int_0^{s^*} 2s P^{-}_{\text{train}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}) \le s\right\} ds\right\} \\
=& \frac{1}{1-\epsilon}\int_0^{s^*} s^2\, dP^{-}_{\text{train}}\left\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x}) \le s\right\}
\end{aligned} \tag{41}$$

Thus, we obtain the following upper bound on the outlier robust risk $\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta}; P^{-}_{\text{train}}, \delta, \epsilon)$ given in (36)

$$\begin{aligned}
\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta}; P^{-}_{\text{train}}, \delta, \epsilon) \le& \mathbb{E}_{P^*}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\big\} + \sqrt{\delta \mathbb{V}_{P^*}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\big\}} \\
=& \Re_1 + \sqrt{\delta(\Re_2 - \Re_1^2)},
\end{aligned}$$

where $\Re_1$ and $\Re_2$ are given in (40) and (41), respectively.

Case 2: KL-divergence. If the discrepancy metric d in (3) is chosen as the KL-divergence, by (4) and Lemma 5, we have that

$$\begin{aligned}
\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta}; P^{-}_{\text{train}}, \delta, \epsilon) =& \inf_{P' \in \mathcal{P}_p(\mathcal{X})} \left\{\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta}; P', \delta) : \exists \widetilde{P}' \in \mathcal{P}_p(\mathcal{X})\ s.t.\ P^{-}_{\text{train}} = (1-\epsilon)P' + \epsilon \widetilde{P}'\right\} \\
=& \inf_{P' \in \mathcal{P}_p(\mathcal{X})} \Big\{\mathbb{E}_{P'}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\big\} + \sqrt{2\delta \mathbb{V}_{P'}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-};\mathbf{x})\big\}} : \\
& \qquad \exists \widetilde{P}' \in \mathcal{P}_p(\mathcal{X})\ s.t.\ P^{-}_{\text{train}} = (1-\epsilon)P' + \epsilon \widetilde{P}'\Big\}
\end{aligned}$$

Similar to the proof in Case 1 with χ^2-divergence, we construct the distribution P^* in (39) and obtain the following upper bound on the outlier robust risk $\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta}; P^{-}_{\text{train}}, \delta, \epsilon)$:

$$\begin{aligned}\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta}; P^{-}_{\text{train}}, \delta, \epsilon) \leq& \mathbb{E}_{P^*}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\big\} + \sqrt{2\delta \mathbb{V}_{P^*}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\big\}} \\ =& \Re_1 + \sqrt{2\delta(\Re_2 - \Re_1^2)},\end{aligned}$$

where $\Re_1$ and $\Re_2$ are given in (40) and (41), respectively.

Case 3:p-Wasserstein distance. If the discrepancy metric d in (3) is chosen as the p-Wasserstein distance, by (4) and Lemma 5, we have that

$$\begin{aligned}\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta}; P^{-}_{\text{train}}, \delta, \epsilon) &= \inf_{P' \in \mathcal{P}_p(\mathcal{X})} \Big\{\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta}; P', \delta) : \exists \widetilde{P}' \in \mathcal{P}_p(\mathcal{X}) \ s.t. \ P^{-}_{\text{train}} = (1-\epsilon)P' + \epsilon \widetilde{P}'\Big\}. \\ &\leq \inf_{P' \in \mathcal{P}_p(\mathcal{X})} \Big\{\mathbb{E}_{P'}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\big\} + \delta \left\{\mathbb{E}_{P'}\|\nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\|_*^q\right\}^{1/q} + \mathcal{O}(\delta^{2\wedge p}) : \\ &\qquad \exists \widetilde{P}' \in \mathcal{P}_p(\mathcal{X}) \ s.t. \ P^{-}_{\text{train}} = (1-\epsilon)P' + \epsilon \widetilde{P}'\Big\}\end{aligned}$$

Similar to the proof in Case 1 with χ^2-divergence, we construct the distribution P^* in (39) and obtain the following upper bound on the outlier robust risk $\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta}; P^{-}_{\text{train}}, \delta, \epsilon)$:

$$\begin{aligned}\mathcal{R}^{-}_{\mathbf{x}}(\boldsymbol{\theta}; P^{-}_{\text{train}}, \delta, \epsilon) \leq& \mathbb{E}_{P^*}\big\{\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\big\} + \delta \left\{\mathbb{E}_{P^*}\|\nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\|_*^q\right\}^{1/q} \\ =& \Re_1 + \delta \left\{\mathbb{E}_{P^*}\|\nabla \mathcal{S}_{\boldsymbol{\theta}}(\mathbf{X}^{-}; \mathbf{x})\|_*^q\right\}^{1/q},\end{aligned}$$

where $\Re_1$ is given in (40).

□

A.4 Proof of Theorem 4.2

Proof of Theorem 4.2. By Lemma 1, when the p-Wasserstein distance with $0-1$ cost is used to construct the uncertainty set, the minimax problem (5) can be equivalently expressed as:

$$\inf_{p \in \Delta^{K-1}} \inf_{\gamma \geq 0} \left[\gamma\delta^p + \sum_{j=1}^{K} p_j^{+} \max\{-p_1 - \gamma, \ldots, -p_j, \ldots, 1 - p_K - \gamma\}\right]. \tag{42}$$

Additionally, we denote

$$g(\gamma; p) \triangleq \gamma\delta^p + \sum_{j=1}^{K} p_j^{+} \max\{-p_1 - \gamma, \ldots, -p_j, \ldots, -p_K - \gamma\}. \tag{43}$$

Step 1: Optimal Lagrange multiplier. We first consider the optimal Lagrange multiplier, denoted $\gamma^\star$, for *each fixed* p. For a fixed p, we sort $\{p_1, \ldots, p_K\}$ in an decreasing order, denoted $p_{(1)} \geq \ldots \geq p_{(K)}$, and hence, $1 - p_{(1)} \leq \ldots \leq 1 - p_{(K)}$. Assume that $\{p_{(1)}, \ldots, p_{(K)}\}$ corresponds to $\{p_1, \ldots, p_K\}$ via a permutation χ, that is, $p_{(j)} = p_{\chi(j)}$. And correspondingly, the p_j^{+}'s with the associated indexes are denoted $p_{(j)}^{+} \triangleq p_{\chi(j)}^{+}$ for $j \in [K]$.

If $1 - p_{(1)} \geq 1 - p_{(K)} - \gamma\kappa^p$, i.e., $\gamma \geq p_{(1)} - p_{(K)}$, by (43), we then obtain that

$$g(\gamma; p) = \gamma\delta^p + \sum_{j=1}^{K} p_j^{+}(1 - p_j) \tag{44}$$

is increasing in γ. Hence, it suffices to consider the case $0 \leq \gamma \leq p_{(1)} - p_{(K)}$.

For $s = 1, 2, \ldots, K-1$, if $1 - p_{(s)} < 1 - p_{(K)} - \gamma\kappa^p \le 1 - p_{(s+1)}$, i.e., $p_{(s+1)} - p_{(K)} \le \gamma < p_{(s)} - p_{(K)}$, we have that

$$
\begin{aligned}
g(\gamma; p) &= \gamma\delta^p + \sum_{j=1}^{s} p^+_{(j)}(1 - p_{(K)} - \gamma\kappa^p) + \sum_{j=s+1}^{K} p^+_{(j)}(1 - p_{(j)}) \\
&= \sum_{j=1}^{s} p^+_{(j)}(1 - p_{(K)}) + \sum_{j=s+1}^{K} p^+_{(j)}(1 - p_{(j)}) + \gamma\kappa^p\Big\{\delta^p - \sum_{j=1}^{s} p^+_{(j)}\Big\}.
\end{aligned} \tag{45}
$$

As

$$
\begin{aligned}
&\lim_{\gamma \to ((p_{(s)} - p_{(K)})/\kappa^p)-} g(\gamma; p) \\
=&\sum_{j=1}^{s} p^+_{(j)}(1 - p_{(K)}) + \sum_{j=s+1}^{K} p^+_{(j)}(1 - p_{(j)}) + \big(p_{(s)} - p_{(K)}\big)\Big\{\delta^p - \sum_{j=1}^{s} p^+_{(j)}\Big\} \\
=&\sum_{j=1}^{s} p^+_{(j)}(1 - p_{(K)}) + \sum_{j=s+1}^{K} p^+_{(j)}(1 - p_{(j)}) + \big(p_{(s)} - p_{(K)}\big)\Big\{\delta^p - \sum_{j=1}^{s-1} p^+_{(j)}\Big\} \\
&- p^+_{(s)}\big\{(1 - p_{(K)}) - (1 - p_{(s)})\big\} \\
=&\sum_{j=1}^{s-1} p^+_{(j)}(1 - p_{(K)}) + \sum_{j=s}^{K} p^+_{(j)}(1 - p_{(j)}) + \big(p_{(s)} - p_{(K)}\big)\Big\{\delta^p - \sum_{j=1}^{s-1} p^+_{(j)}\Big\} \\
=&g((p_{(s)} - p_{(K)})/\kappa^p; p),
\end{aligned}
$$

we have that $g(\gamma; p)$ is continuous in γ for $0 \le \gamma \le p^K - p^1$.

If $p^+_{(1)} < \delta^p < \sum_{j=1}^{K} p^+_{(j)}$, then there exists an $s^* \in \{2, \ldots, K\}$ such that $\sum_{j=1}^{s^*-1} p^+_{(j)} \le \delta^p \le \sum_{j=1}^{s^*} p^+_{(j)}$. Then, by (45), we obtain that $g(\gamma; p)$ is decreasing in γ for $\gamma \in [0, p_{(s^*)} - p_{(K)}]$ and increasing for $\gamma \in [p_{(s^*)} - p_{(K)}, p_{(1)} - p_{(K)}]$. If $\delta^p \le p^+_{(1)}$, let $s^* = 1$, and $g(\gamma; p)$ is decreasing in γ for $\gamma \in [0, p_{(s^*)} - p_{(K)}]$; if $\delta^p \ge \sum_{j=1}^{K} p^+_{(j)}$, let $s^* = K$, and $g(\gamma; p)$ is increasing in γ for $\gamma \in [p_{(s^*)} - p_{(K)}, p_{(1)} - p_{(K)}]$. Hence, the optimal Lagrange multiplier is given as $\gamma^\star \triangleq p_{(s^*)} - p_{(K)}$.

Step 2: Linear programming format. For *each fixed permutation* χ, we next show the format of the optimal p that minimizes $g(\gamma^\star; p)$.

If $s^* = K$, then $g(\gamma^\star; p) = g(0; p) = 1 - p_{(K)} \ge 1 - 1/K$, and the corresponding optimal action is $p_{(1)} = \ldots = p_{(K)} = 1/K$.

If $s^* \in [K-1]$, by (45), and the robust risk for a single data point $(\mathbf{x}, \widetilde{\mathbf{y}})$ is computed as

$$
\begin{aligned}
g(\gamma^\star; p) &= \sum_{j=1}^{s^*-1} p^+_{(j)}(1 - p_{(K)})\mathbf{1}(s^* > 1) + \sum_{j=s^*}^{K} p^+_{(j)}(1 - p_{(j)}) \\
&\quad + (p_{(s^*)} - p_{(K)})\Big\{\delta^p - \sum_{j=1}^{s^*-1} p^+_{(j)}\mathbf{1}(s^* > 1)\Big\} \\
&=\Big\{\sum_{j=1}^{s^*} p^+_{(j)} - \delta^p\Big\}(1 - p_{(s^*)}) + \sum_{j=s^*+1}^{K-1} p^+_{(j)}(1 - p_{(j)})\mathbf{1}(s^* < K-1) \\
&\quad + \Big\{p^+_{(K)} + \delta^p\Big\}(1 - p_{(K)}).
\end{aligned}
$$

Let $z_j \triangleq 1 - p_{(j)}$. Then, the optimal p can be derived by solving the following linear programming problem:

$$\begin{cases} \min\limits_{z_1,\ldots,z_K} \mathsf{V}(\mathbf{z}) = \sum\limits_{j=s^*}^{K} a_j z_j \\ s.t. \sum\limits_{j=1}^{K}(1-z_j) = 1, \\ 0 \le z_1 \le \ldots \le z_K \le 1, \end{cases} \tag{46}$$

where $\mathsf{V}(\mathbf{z})$ is called the value function at $\mathbf{z}$, and a_j's are the corresponding nonnegative coefficients. For each $\mathbf{z} = (z_1, \ldots, z_{s^*}, z_{s^*+1}, \ldots, z_K)^\top$ in the feasible region of (46), denoted Ξ, let $\widetilde{\mathbf{z}} = (z_{s^*}, \ldots, z_{s^*}, \widetilde{z}_{s^*+1}, \ldots, \widetilde{z}_K)^\top$, where $\widetilde{z}_j = z_j - c$ for $j = s^*+1, \ldots, K$, with $c \triangleq (s^* \cdot z_{s^*} - \sum_{j=1}^{s^*} z_j)/(K - s^*) \ge 0$. Then, $\widetilde{\mathbf{z}} \in \Xi$ and $\mathsf{V}(\widetilde{\mathbf{z}}) \le \mathsf{V}(\mathbf{z})$ by the nonnegativity of a_j's. Therefore, we can only consider the optimal values of $\{z_{s^*}, \ldots, z_K\}$, and the linear programming problem (46) can be equivalently written as below:

$$\begin{cases} \min\limits_{z_{s^*},\ldots,z_K} \mathsf{V}(\mathbf{z}) = \sum\limits_{j=s^*}^{K} a_j z_j \\ s.t.\ s^* \cdot (1 - z_{s^*}) + \sum\limits_{j=s^*+1}^{K}(1-z_j) = 1, \\ 0 \le z_{s^*} \le \ldots \le z_K \le 1. \end{cases} \tag{47}$$

Moreover, the feasible region of (47), denoted $\overline{\Xi}$, can also be expressed as follows:

$$\begin{aligned} \overline{\Xi} \triangleq & \Big\{ z_{s^*}, \ldots, z_K : s^* \cdot (1 - z_{s^*}) + \sum_{j=s^*+1}^{K}(1-z_j) = 1,\ 0 \le z_{s^*} \le \ldots \le z_K \le 1, \\ & \qquad 1 - z_K \le \frac{1}{K}, 1 - z_{K-1} \le \frac{1}{K-1}, \ldots, 1 - z_{s^*} \le \frac{1}{s^*} \Big\} \\ = & \Big\{ z_{s^*}, \ldots, z_K : s^* \cdot z_{s^*} + \sum_{j=s^*+1}^{K} z_j = K - 1,\ 0 \le z_{s^*} \le \ldots \le z_K \le 1, \\ & \qquad z_K \ge 1 - \frac{1}{K}, z_{K-1} \ge 1 - \frac{1}{K-1}, \ldots, z_{s^*} \ge 1 - \frac{1}{s^*} \Big\}. \end{aligned}$$

Step 3: Extreme points. We next prove that the following $K - s^* + 1$ feasible solutions are the only extreme points of (47):

$$\begin{aligned} & \mathbf{z}_1 \triangleq (1 - \frac{1}{s^*}, 1, 1, \ldots, 1, 1)^\top, \\ & \mathbf{z}_2 \triangleq (1 - \frac{1}{s^*+1}, 1 - \frac{1}{s^*+1}, 1, \ldots, 1, 1)^\top, \\ & \ldots, \\ & \mathbf{z}_j = (1 - \frac{1}{s^*+j-1}, \ldots, 1 - \frac{1}{s^*+j-1}, 1, \ldots, 1)^\top \\ & \ldots, \\ & \mathbf{z}_{K-s^*} \triangleq (1 - \frac{1}{K-1}, 1 - \frac{1}{K-1}, 1 - \frac{1}{K-1}, \ldots, 1 - \frac{1}{K-1}, 1)^\top, \\ & \mathbf{z}_{K-s^*+1} \triangleq (1 - \frac{1}{K}, 1 - \frac{1}{K}, 1 - \frac{1}{K}, \ldots, 1 - \frac{1}{K}, 1 - \frac{1}{K})^\top. \end{aligned}$$

We denote $\overline{\Theta}_0 \triangleq \{\mathbf{z}_1, \ldots, \mathbf{z}_{K-s^*+1}\}$.

Firstly, we prove that each data point in $\overline{\Theta}_0$ is an extreme point of (47). In particular, for $j \in [K - s^* + 1]$, suppose that $\mathbf{z}_j = \nu \mathbf{z}' + (1-\nu)\mathbf{z}''$ for some $\nu \in (0, 1)$, with $\mathbf{z}' = (z'_{s^*}, \ldots, z'_K)^\top \in \overline{\Xi}$ and

$\mathbf{z}'' = (z''_{s^*}, \ldots, z''_K)^\top \in \overline{\Xi}$. For $t = s^* + j, \ldots, K$, since $\nu z'_t + (1-\nu) z''_t = z_{j,t} = 1$ and $z'_t, z''_t \le 1$, we have that $z'_t = z''_t = z_{j,t}$ with $z_{j,t}$ denoting the tth element of $\mathbf{z}_j$. Additionally, we obtain that $z'_{s^*+j-1} = z''_{s^*+j-1} = z_{j,s^*+j-1}$ as $\nu z'_{s^*+j-1} + (1-\nu) z''_{s^*+j-1} = z_{j,s^*+j-1} = 1 - \frac{1}{s^*+j-1}$ and $z'_{s^*+j-1}, z''_{s^*+j-1} \ge 1 - \frac{1}{s^*+j-1}$. Moreover, for $t < s^* + j - 1$, since $z'_t \le z'_{s^*+j-1} = 1 - \frac{1}{s^*+j-1}$, $z''_t \le z''_{s^*+j-1} = 1 - \frac{1}{s^*+j-1}$, and $\nu z'_t + (1-\nu) z''_t = z_{j,t} = 1 - \frac{1}{s^*+j-1}$, we can also obtain that $z'_t = z''_t = z_{j,t}$. Therefore, $\mathbf{z}' = \mathbf{z}'' = \mathbf{z}_j$, and hence, $\mathbf{z}_j$ is an extreme point of (47) by Definition A.3.

We next consider a point $\widetilde{\mathbf{z}} \triangleq (\widetilde{z}_{s^*}, \ldots, \widetilde{z}_K)^\top \in \overline{\Theta} \backslash \overline{\Theta}_0$ and prove it is not an extreme point of (47) by construction. Specifically, we have the following claims for $\widetilde{\mathbf{z}}$.

- $\widetilde{z}_t > 1 - \frac{1}{t}$ for $t = s^*, \ldots, K$.
  - This claim can be proved by contradiction. If there exists $t_0 \in \{s^*, \ldots, K\}$ such that $\widetilde{z}_{t_0} = 1 - \frac{1}{t_0}$, we have that $s^* \cdot \widetilde{z}_{s^*} + \sum_{j=s^*+1}^{K} z_j \le t_0 \cdot \widetilde{z}_{t_0} + (K - t_0) \cdot 1 = K - 1$. Here the inequality holds if and only if $\widetilde{z}_t = \widetilde{z}_{t_0} = 1 - \frac{1}{t_0}$ for $t < t_0$ and $\widetilde{z}_t = 1$ for $t > t_0$; that is, in this case, $\widetilde{\mathbf{z}}$ falls in the feasible region $\overline{\Xi}$ if and only if $\widetilde{\mathbf{z}}$ is one of the aforementioned $K - s^* + 1$ extreme points.
- There exists $t_1 \in \{s^* + 1, \ldots, K\}$ such that $\widetilde{z}_{t_1 - 1} < \widetilde{z}_{t_1} < 1$. Let $t_2 \triangleq \max_{t \in \{s^*+1, \ldots, K\}} \{\widetilde{z}_t < 1\}$. Then, $t_2 \ge t_1$.
  - This claim can be proved by contradiction. In particular, we assume the claim is not true. If there exists $t_1 \in \{s^* + 1, \ldots, K\}$ such that $\widetilde{z}_{t_1-1} < \widetilde{z}_{t_1}$, then we have $\widetilde{z}_t = \widetilde{z}_{t_1 - 1}$ for $t \le t_1 - 1$ and $\widetilde{z}_t = 1$ for $t \ge t_1$ by assumption, and hence, $\widetilde{\mathbf{z}} \in \overline{\Theta}_0$. If $\widetilde{z}_{t-1} = \widetilde{z}_t$ for all $t \in \{s^* + 1, \ldots, K\}$, then $\widetilde{\mathbf{z}} = \mathbf{z}_{K - s^* + 1} \in \overline{\Theta}_0$.

Let $\not{c}_1 \triangleq \min\{\frac{\widetilde{z}_{t_1} - \widetilde{z}_{t_1-1}}{2}, \widetilde{z}_t - (1 - \frac{1}{t})$ for $s^* \le t \le t_1 - 1\}$, $\not{c}_2 \triangleq \min\{\frac{\widetilde{z}_{t_1} - \widetilde{z}_{t_1-1}}{2}, \widetilde{z}_{t_1} - (1 - \frac{1}{t_1}), 1 - \widetilde{z}_t$ for $t_1 \le t \le t_2\}$, $\overline{\not{c}} \triangleq \min\{(t_1 - 1)\not{c}_1, (t_2 - t_1 + 1)\not{c}_2\}$, $\overline{\not{c}}_1 \triangleq \overline{\not{c}}/(t_1 - 1)$, and $\overline{\not{c}}_2 \triangleq \overline{\not{c}}/(t_2 - t_1 + 1)$. Then we construct two points in $\overline{\Theta}$: $\mathbf{z}' \triangleq (\widetilde{z}_{s^*} + \overline{\not{c}}_1, \ldots, \widetilde{z}_{t_1-1} + \overline{\not{c}}_1, \widetilde{z}_{t_1} - \overline{\not{c}}_2, \ldots, \widetilde{z}_{t_2} - \overline{\not{c}}_2, \ldots, \widetilde{z}_K)^\top$, and $\mathbf{z}'' \triangleq (\widetilde{z}_{s^*} - \overline{\not{c}}_1, \ldots, \widetilde{z}_{t_1-1} - \overline{\not{c}}_1, \widetilde{z}_{t_1} + \overline{\not{c}}_2, \ldots, \widetilde{z}_{t_2} + \overline{\not{c}}_2, \ldots, \widetilde{z}_K)^\top$. Therefore, $\widetilde{\mathbf{z}} = \frac{1}{2}\mathbf{z}' + \frac{1}{2}\mathbf{z}''$, and hence, $\widetilde{\mathbf{z}}$ is not an extreme point of (47).

Step 4: Solution format and optimal action. By Step 2 and Step 3, we obtain that *for each fixed χ and s^**, the extreme points of the linear programming problem are given in $\overline{\Theta}_0$. By Lemma 4, every linear program has an extreme point that is an optimal solution. Hence, we obtain that at least one optimal action of p can be found in the format:

$$p_{(j)} = \frac{1}{k^*} \text{ for } j \le k^* \text{ and } p_{(j)} = 0 \text{ for } j \ge k^* \tag{48}$$

for some $k^* \in [K]$.

If $k^* = K$, by (43), we have that $g(\gamma; p) = \gamma \delta^p + \sum_{j=1}^{K} p_j^+ \cdot (1 - \frac{1}{K})$, and hence, the robust risk is $g(\gamma^\star; p) = 1 - \frac{1}{K}$ by taking $\gamma^\star = 0$. If $k^* < K$, we obtain that

$$\begin{aligned} g(\gamma; p) =& \gamma \delta^p + \sum_{j=1}^{k^*} p_{(j)}^+ \max\{1 - \gamma \kappa^p, 1 - \frac{1}{k^*}\} + \sum_{j=k^*+1}^{K} p_{(j)}^+ \cdot 1 \\ =& \begin{cases} 1 + \gamma \kappa^p \Big\{ \delta^p - \sum_{j=1}^{k^*} p_{(j)}^+ \Big\}, \text{ if } 0 \le \gamma \le \dfrac{1}{k^* \kappa^p}; \\ \gamma \delta^p + 1 - \dfrac{1}{k^*} \sum_{j=1}^{k^*} p_{(j)}^+, \text{ if } \gamma \ge \dfrac{1}{k^* \kappa^p}. \end{cases} \end{aligned}$$

Hence, for $k^* < K$, the robust risk is the minimum of $g(\gamma^\star; p) = 1$ by taking $\gamma^\star = 0$ and $g(\gamma^\star; p) = 1 + \frac{1}{k^*}\Big\{\delta^p - \sum_{j=1}^{k^*} p_{(j)}^+\Big\}$ by taking $\gamma^\star = \frac{1}{k^* \kappa^p}$. Additionally, we observe that we should take the highest k^* values of $\{p_1^+, \ldots, p_K^+\}$ as $p_{(1)}^+, \ldots, p_{(k^*)}^+$ to minimize the robust risk. Hence, we take the permutation χ such that $p_{(1)}^+ \geq \ldots \geq p_{(K)}^+$.

In summary, the optimal action $p^\star$ is given as below.

- If $\frac{1}{K} \geq \frac{1}{k^*} \sum_{j=1}^{k^*} p_{(j)}^+ - \frac{1}{k^*}\delta^p$ for all $k^* \in [K-1]$, then $p_j^\star = \frac{1}{K}$ for $j \in [K]$.
- If there exists some $k_0 \in [K-1]$, $\frac{1}{k_0} \sum_{j=1}^{k_0} p_{(j)}^+ - \frac{1}{k_0}\delta^p > \frac{1}{K}$, and $\frac{1}{k_0} \sum_{j=1}^{k_0} p_{(j)}^+ - \frac{1}{k_0}\delta^p \geq \frac{1}{k^*} \sum_{j=1}^{k^*} p_{(j)}^+ - \frac{1}{k^*}\delta^p$ for all $k^* \in [K-1]$, then $p^{\star(j)} = \frac{1}{k_0}$ for $j \in [k_0]$ and $p^{\star(j)} = 0$ for $j = k_0 + 1, \ldots, K$.

In particular, if $p_{(1)}^+ \geq \max\{\frac{1}{K} + \delta^p, p_{(2)}^+ + \delta^p\}$, then the optimal action is given as: $p^{\star(1)} = 1$ and $p^{\star(j)} = 0$ for $j = 2, \ldots, K$. Thus, the proof is complete. $\square$

B Experimental Details

Source Models. For the source models, we use those provided by Liang et al. (2020) and Yang et al. (2021a) for the Office-Home and VisDA2017 datasets. Since no open-source models were available for Office-31 and DomainNet-126, we trained the source models ourselves using the training methodologies from SHOT (Liang et al., 2020) and C-SFDA (Karim et al., 2023), respectively.

Target Adaptation Training. We train both the model backbone and classifier during the adaptation process, primarily following the SHOT (Liang et al., 2020) and AaD (Yang et al., 2022) setup. For the optimizer, we use SGD with momentum of 0.9 and weight decay of $1e^{-3}$. We also use the Nesterov update method. The initial learning rate for the bottleneck and classification layers is set to 0.001 across all datasets. For the backbone models, the initial learning rates are set as follows: $5e^{-4}$ for Office-Home, $1e^{-4}$ for DomainNet-126 and Office-31, and $5e^{-5}$ for VisDA2017. We use the same learning rate scheduler as Liang et al. (2020) for the Office-Home and DomainNet-126 datasets. The batch size is 64 for all datasets. We train for 30 epochs on VisDA2017 and 45 epochs on Office-Home, Office-31, and DomainNet-126. All experiments are run on a single 32GB V100 or 40GB A100 GPU.

Hyperparameters Selection. In SFDA, hyperparameter selection presents a significant challenge due to the lack of labeled target data and the distribution shift between domains. In our experiments, we followed the common pipeline for hyperparameter tuning in the literature (e.g., Yang et al. (2022); Hwang et al. (2024)), and employed the SND (Soft Neighborhood Density) score (Saito et al., 2021) and sensitivity analysis to guide the hyperparameter selection.

In fact, most hyperparameters in our method do not require intensive tuning, and their selection can be guided by our theoretical analysis outlined below.

Our UCon-SFDA method consists of three main components: the basic contrastive loss $\mathcal{L}_{\mathrm{CL}}$, the dispersion control term $\mathcal{L}^{-}_{\mathrm{DC}}$, and the partial label term $\mathcal{L}^{+}_{\mathrm{PL}}$. Given the complexity of the parameter space, we simplified the hyperparameter selection process by avoiding exhaustive consideration of all parameter combinations. Instead, we adopted a **sequential**, **incremental** approach to tune the parameters for the three loss terms, one at a time.

First, for the hyperparameters in the $\mathcal{L}_{\mathrm{CL}}$ terms (first three columns in Table 6), including the number of positive samples κ, the decay exponent β for the negative term, and the negative sample loss coefficient $\lambda^{-}_{\mathrm{CL}}$, we largely follow the configurations used in Yang et al. (2022) and Hwang et al. (2024). As in previous works, we directly set $\lambda^{-}_{\mathrm{CL}}$ to 1. For datasets with more classification categories, such as Office-Home, Office, and DomainNet-126, where noise in negative samples is less pronounced, we use a smaller decay exponent to enhance the impact of true-negative samples during adaptation. In contrast, for VisDA, which contains only 12 classes with a batch size of 64, we apply a faster decay rate to mitigate the influence of false-negative samples.

Next, we consider the hyperparameter associated with the dispersion term, λ_{DC}. In our initial experimental trials, we set this value to either 0.5 or 1, based on a balance between the loss terms, $\mathcal{L}^{+}_{\mathrm{CL}}$ and $\mathcal{L}^{-}_{\mathrm{DC}}$, and the sensitivity analysis of hyperparameters.

Finally, for the hyperparameters λ_{PL}, K_{PL}, and τ in the partial label loss, we also performed the basic sequential tuning under the guidance of theoretical insights. According to the proposed algorithm, we use τ to select the uncertain data points and merge the top-K_{PL} predicted classes into the partial label set for each selected data point. Theoretically, a smaller τ (yet naturally larger than 1) represents a more uncertain set. As we want to apply the partial label loss only on the uncertain data points and avoid the introduction of additional label uncertainty for more confident data points, we considered a value in $\{1.1, 1.3, 1.5\}$ for τ. We found that $\tau = 1.1$ is sufficient for achieving promising performance, except for simpler tasks with high initial prediction accuracy, such as Office-31. Next, the value of the partial label number K_{PL} should be determined based on the algorithm and the number of categories in the dataset. Generally, a small K_{PL} is preferred, as the partial label set is gradually enlarged with each epoch. A large K_{PL} could result in an overly large partial label set, potentially introducing more uncertainty. Empirically, we evaluated $K_{\mathrm{PL}} \in \{1, 2, 3\}$, and found that $K_{\mathrm{PL}} = 2$ performs well for most datasets, except for VisDA2017, whose total number of classes is only 12 and $K_{\mathrm{PL}} = 1$ is sufficient. Finally, we tuned λ_{PL} by considering $\lambda_{\mathrm{PL}} \in \{0.001, 0.01, 0.05, 0.1\}$ and selected the best-performing value based on the guidance of the hyperparameter sensitivity analyses.

The final selected parameter values used in our experiments are summarized in Table 6, which are obtained by a relatively straightforward tuning process conducted on a subspace of hyperparameters. We note that more refined tuning over the full combinatorial hyperparameter space can further enhance the performance of our algorithm; additional analysis on the sensitivity of these hyperparameters is provided in Appendix C.5.

Table 6: Hypermaraters on Different Datasets.

Dataset	κ	$\lambda_{\mathrm{CL}}^{-}$	β	λ_{DC}	λ_{PL}	K_{PL}	τ
Office-31	3	1	1	1	0.05	2	1.3
Office-Home	3	1	0	0.5	0.001	2	1.1
Office-Home (partial set)	5	1	0.75	1	0.1	2	1.1
VisDA2017	5	1	5	1	0.01	1	1.1
VisDA-RUST	3	1	5	0.5	0.1	2	1.1
DomainNet-126	2	1	0.75	0.5	0.1	2	1.1

Algorithm. The overall description of adaptation process with our UCon-SFDA method is shown in Algorithm 1

Algorithm 1: UCon-SFDA - Uncertainty-Controlled Source-Free Domain Adaptation

Input: Pre-Trained Source Model: $f_{\mathrm{S}}(\mathbf{x};\boldsymbol{\theta})$, Target Data: $\mathcal{D}_{\mathrm{T}} \triangleq \{\mathbf{x}_i^{\mathrm{T}}\}_{i=1}^{N_{\mathrm{T}}}$, Training Epochs: T,

```
// Initialization Process
```
Initialize a target model $f_{\mathrm{T}}(\mathbf{x};\boldsymbol{\theta}_0) = f_{\mathrm{S}}(\mathbf{x};\boldsymbol{\theta})$

Construct feature bank $\mathscr{Z}$ and predicted score bank $\mathscr{F}$ as described in Yang et al. (2022).

Initialize partial label bank $\mathscr{Y}_{\mathrm{PL}}$ and uncertainty sample bank $\mathscr{U}$ as proposed in Section 4.4.

```
// Training/Adaptation Process
```
for *epoch=1* **to** T **do**

    for *iterations t = 1,2,3,...* **do**

        Forward Propagation: obtain feature $\mathbf{z}_i$, predicted probabilities $f_{\mathrm{T}}(\mathbf{x}_i;\boldsymbol{\theta}_t)$ and $f_{\mathrm{T}}(\mathtt{AUG}(\mathbf{x}_i);\boldsymbol{\theta}_t)$ for each sample $\mathbf{x}_i$ in mini-batch $\mathcal{B}$.

        Bank Refresh: update $\mathscr{Z}$ and $\mathscr{F}$ using $\mathbf{z}_{\mathcal{B}}$ and $f_{\mathrm{T}}(\mathbf{x}_{\mathcal{B}};\boldsymbol{\theta}_t)$ as described in Yang et al. (2022); update $\mathscr{Y}_{\mathrm{PL}}$ and $\mathscr{U}$ as proposed in Section 4.4.

        Compute Negative Uncertainty Control Loss $\mathcal{L}_{\mathrm{UCon}}^{-}$ in Equation (7) using $f_{\mathrm{T}}(\mathbf{x}_{\mathcal{B}};\boldsymbol{\theta}_t)$ and $f_{\mathrm{T}}(\mathtt{AUG}(\mathbf{x}_{\mathcal{B}});\boldsymbol{\theta}_t)$

        Compute Positive Uncertainty Control Loss $\mathcal{L}_{\mathrm{UCon}}^{+}$ in Equation (8) using $\mathscr{Z}$, $\mathscr{F}$, $\mathscr{Y}_{\mathrm{PL}}$ and $\mathscr{U}$.

        Compute the total Uncertainty Control Source-Free Domain Adaptation Loss $\mathcal{L}_{\mathrm{UCon-SFDA}} = \mathcal{L}_{\mathrm{UCon}}^{+} + \mathcal{L}_{\mathrm{UCon}}^{-}$

        Update the parameters of $f_{\mathrm{T}}(\boldsymbol{\theta}_t)$ via $\mathcal{L}_{\mathrm{UCon-SFDA}}$

    end for

end for

Output: Target Adapted Model $f_{\mathrm{T}}(\mathbf{x}_i;\boldsymbol{\theta}_t)$

C Additional Experimental Results

C.1 Experimental Result on office-Home

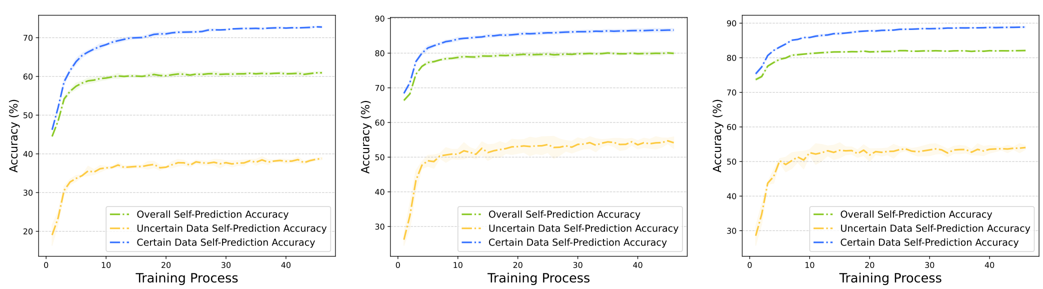
Due to the main text page limitation, we have displayed the experimental result on the Office-Home dataset in the appendix, as shown in Table 7

Table 7: Classification Accuracy (%) on the Office-Home Dataset (ResNet-50)

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., 2020)	57.1	78.1	81.5	68.0	78.2	78.1	67.4	54.9	82.2	73.3	58.8	84.3	71.8
A^{2}Net (Xia et al., 2021)	58.4	79.0	82.4	67.5	79.3	78.9	**68.0**	56.2	82.9	74.1	60.5	85.0	72.8
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
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
CPGA (Qiu et al., 2021)	59.3	78.1	79.8	65.4	75.5	76.4	65.7	58.0	81.0	72.0	**64.4**	83.3	71.6
CoWA-JMDS (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
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
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
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
I-SFDA (Mitsuzumi et al., 2024a)	60.7	78.9	82.0	**69.9**	79.5	79.7	67.1	58.8	82.3	**74.2**	61.3	**86.4**	73.4
UCon-SFDA (Ours)	**61.5**	**80.5**	82.1	69.3	**80.8**	78.7	67.0	**62.2**	82.0	72.2	61.9	85.5	**73.6**

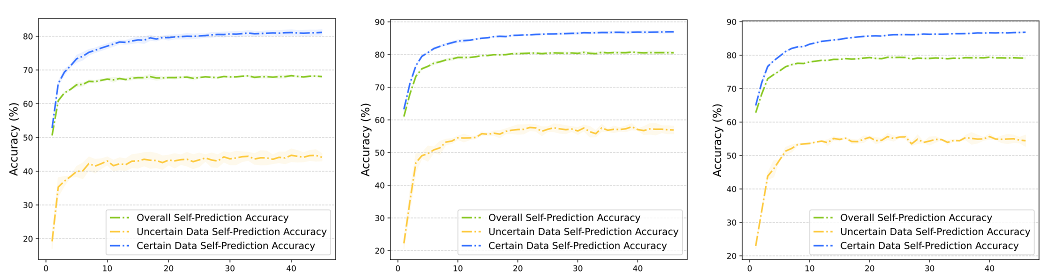
C.2 Partial Label Set Evaluation

We conduct the self-prediction, partial label set, and neighbor label set evaluations across all 12 tasks on the office-home dataset. The results of self-prediction are shown in Figure 4 to Figure 7, and the results of partial label set and neighbor set comparison are shown in Figure 8 to Figure 11



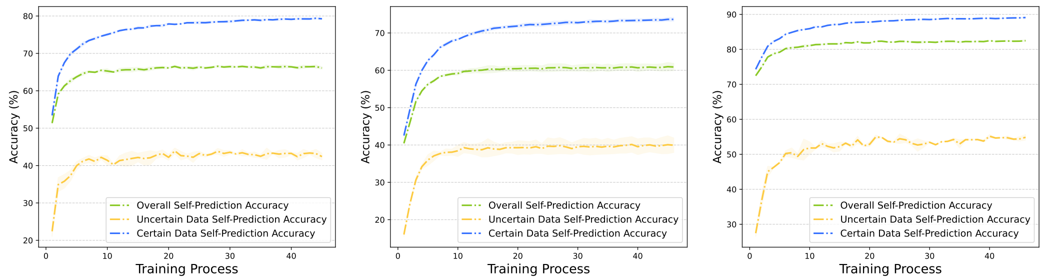
(a) Ar → Cl (b) Ar → Pr (c) Ar → Rw

Figure 4: *Self-prediction accuracy among different data certainty levels on Office-Home Dataset with Source Domain Ar*



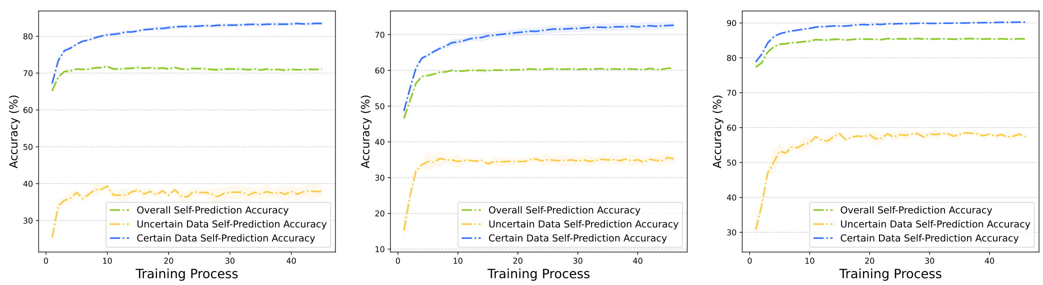
(a) Cl → Ar (b) Cl → Pr (c) Cl → Rw

Figure 5: *Self-prediction accuracy among different data certainty levels on Office-Home Dataset with Source Domain Cl*



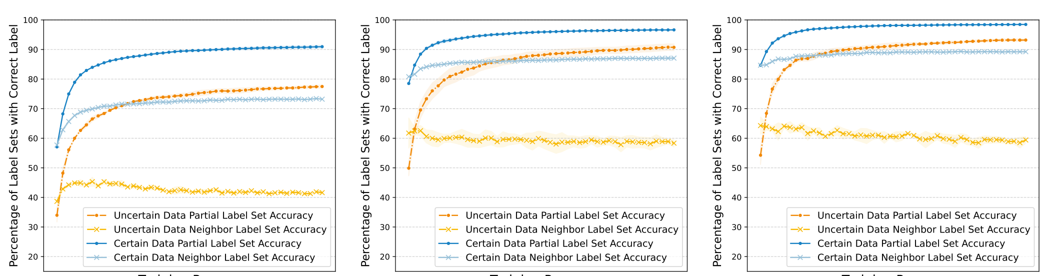
(a) Pr → Ar (b) Pr → Cl (c) Pr → Rw

Figure 6: *Self-prediction accuracy among different data certainty levels on Office-Home Dataset with Source Domain Pr*



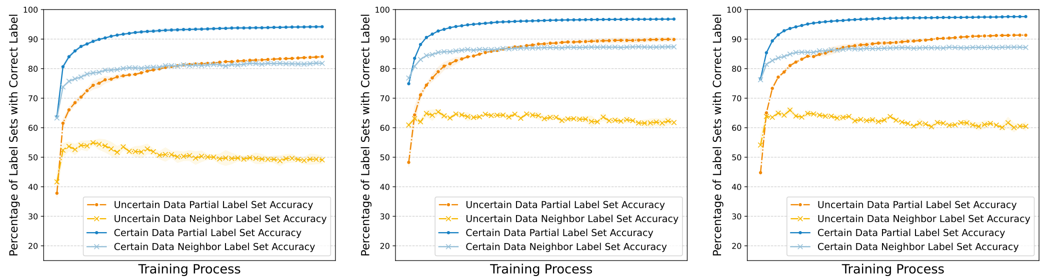
(a) Rw → Ar (b) Rw → Cl (c) Rw → Pr

Figure 7: *Self-prediction accuracy among different data certainty levels on Office-Home Dataset with Source Domain Rw*



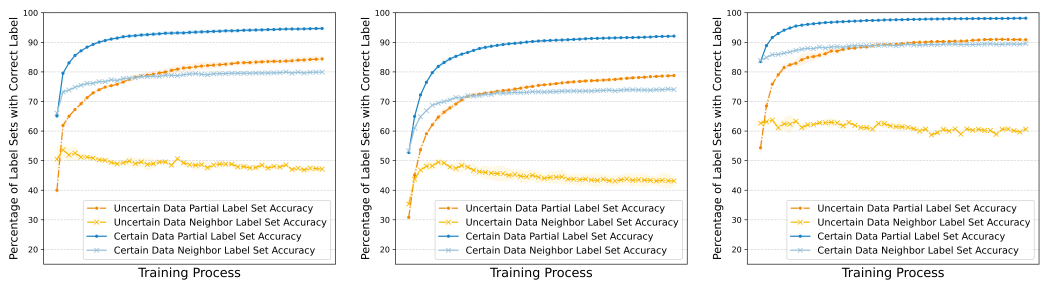
(a) Ar → Cl (b) Ar → Pr (c) Ar → Rw

Figure 8: *Label set Correctness among different data certainty levels on Office-Home Dataset with Source Domain Ar*



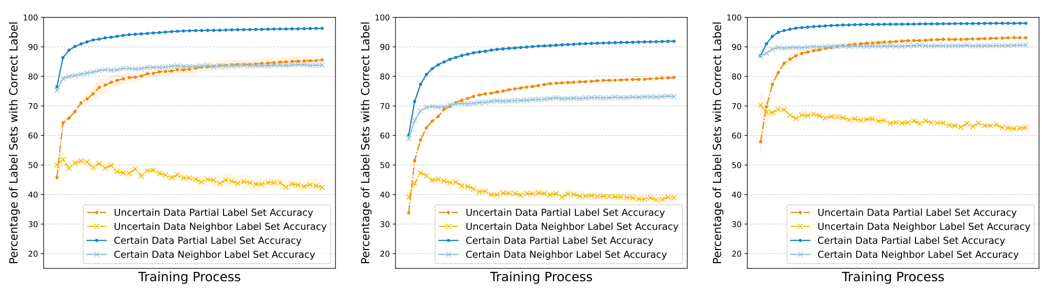
(a) Cl → Ar (b) Cl → Pr (c) Cl → Rw

Figure 9: *Label set Correctness among different data certainty levels on Office-Home Dataset with Source Domain Cl*



(a) Pr → Ar (b) Pr → Cl (c) Pr → Rw

Figure 10: *Label set Correctness among different data certainty levels on Office-Home Dataset with Source Domain Pr*



(a) Rw → Ar (b) Rw → Cl (c) Rw → Pr

Figure 11: *Label set Correctness among different data certainty levels on Office-Home Dataset with Source Domain Rw*

C.3 Data Augmentation in SFDA

We evaluate the prediction accuracies and consistency of original target data and their augmented version by source model on Office-Home and VisDA-2017. The consistency is defined as:

$$\texttt{CONSISTENCY} \triangleq \sum_{i=1}^{N_{\mathrm{T}}} \mathbb{1}_{\{f_{\mathrm{S}}(\mathbf{x}_i;\boldsymbol{\theta})=f_{\mathrm{S}}(\texttt{AUG}(\mathbf{x}_i);\boldsymbol{\theta})\}}.$$

As shown in Figure 12, we can notice that the source model exhibits lower accuracy in predicting the augmented data and demonstrates a high inconsistency between the predictions for the anchor

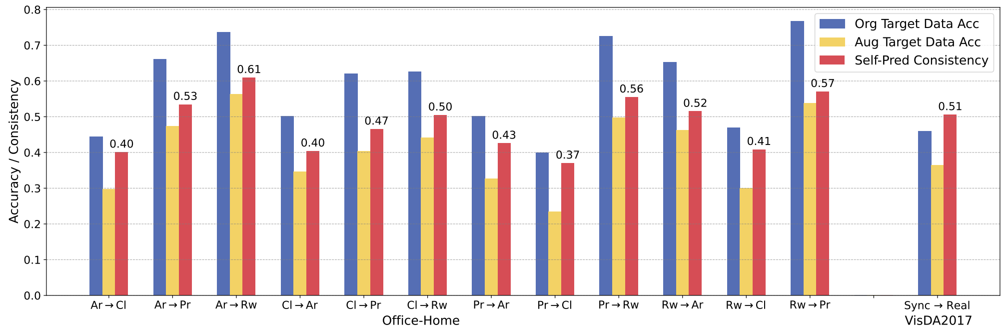
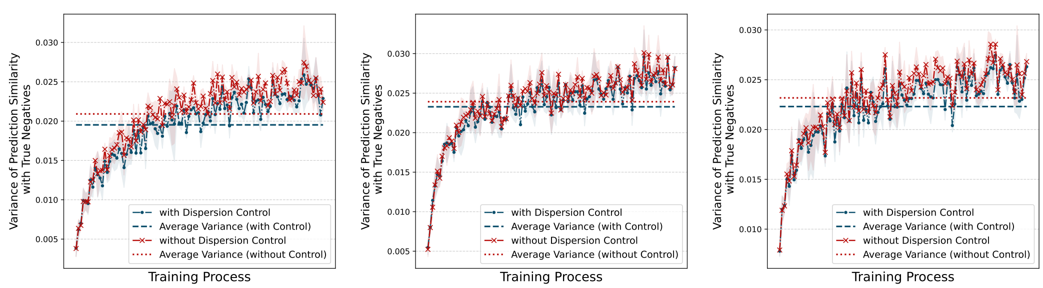


Figure 12: *Inconsistency between the prediction results between the anchor image and its augmented view by source model.*

data and its augmented versions. This experimental result quite contradicts intuitive expectations. It empirically explains why some methods, directly using the augmented predictions as additional labels or supervisory signals, fail to improve SFDA performance effectively, and may even have a negative impact.

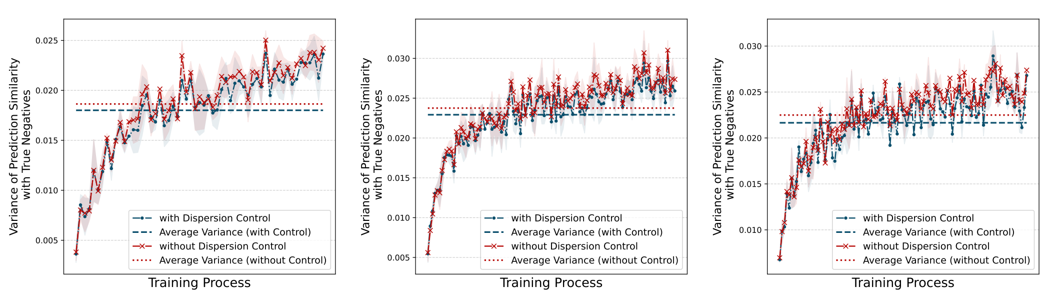
C.4 Variance Control Effect

We evaluate the dispersion control effect achieved by our augmentation-based $\mathcal{L}_{\mathrm{DC}}^{-}$ across all 12 tasks on the office-home dataset. The results are shown in Figure 13 to Figure 16. The consistent dispersion reduction achieved validates the effectiveness of our proposed method.



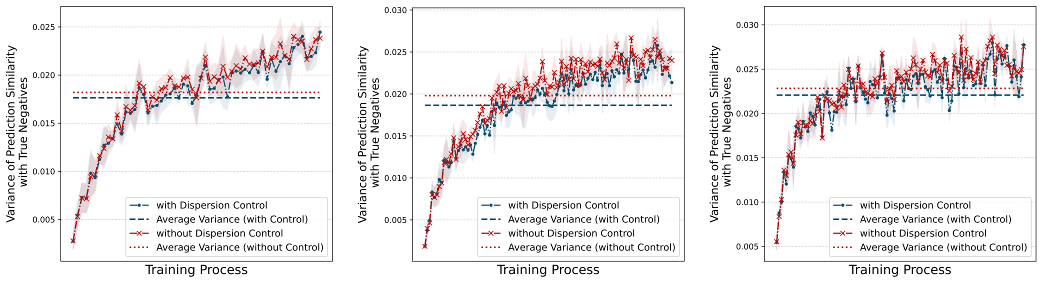
(a) Ar → Cl (b) Ar → Pr (c) Ar → Rw

Figure 13: *Dispersion Control Loss Effect on Office-Home Dataset with Source Domain Ar*



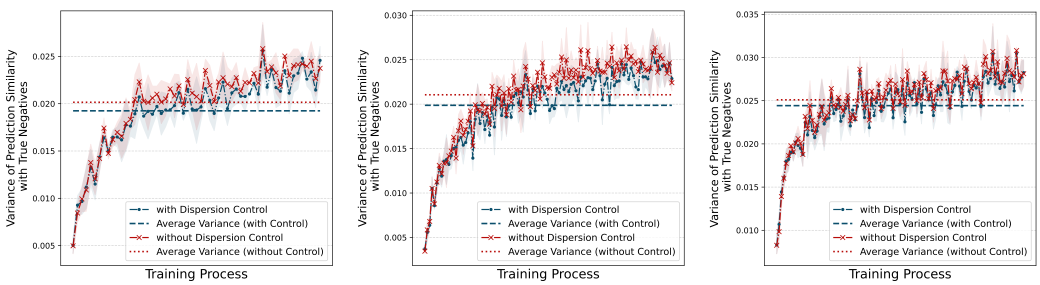
(a) Cl → Ar (b) Cl → Pr (c) Cl → Rw

Figure 14: *Dispersion Control Loss Effect on Office-Home Dataset with Source Domain Cl*



(a) Pr $\rightarrow$ Ar (b) Pr $\rightarrow$ Cl (c) Pr $\rightarrow$ Rw

Figure 15: *Dispersion Control Loss Effect on Office-Home Dataset with Source Domain Pr*



(a) Rw $\rightarrow$ Ar (b) Rw $\rightarrow$ Cl (c) Rw $\rightarrow$ Pr

Figure 16: *Dispersion Control Loss Effect on Office-Home Dataset with Source Domain Rw*

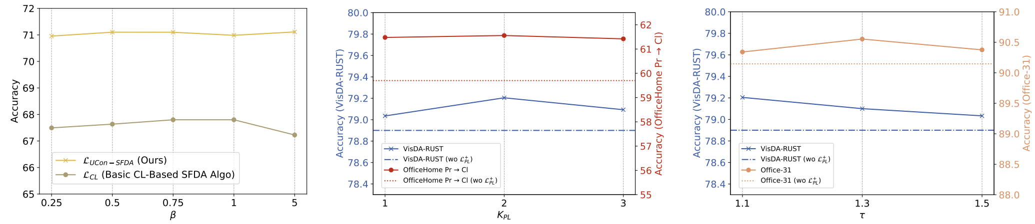
C.5 Sensitivity Analyses of Hyperparameters

To further understand the performance of the proposed method, we conducted comprehensive experiments to study the sensitivity of our method to different choices of hyperparameters involved in our algorithm. While we primarily used the hyperparameter configurations from previous works (Yang et al., 2022; Hwang et al., 2024) for $\lambda^{-}_{\mathrm{CL}}$, κ and β, we also investigated the sensitivity of our method relative to different choices of β, K_{PL}, τ, λ_{PL} and λ_{DC}. The experimental results are summarized in Figure 17(a), (b), (c), Figure 18 and Figure 19, respectively.

Specifically, in Figure 17(a)-(c), the solid lines represent the accuracy of different methods with respect to the different parameter values of β, K_{PL}, and τ. In Figure 17(b)-(c), we added the dashed horizontal lines to indicate the performance on different datasets without the partial label loss for a clear comparison. In Figures 18- 19, the blue, red, and yellow lines represent the accuracy on the target dataset, the accuracy on the small evaluation set, and the SND score, respectively. The shaded regions correspond to the results reported in the main text and the associated parameter values. For Figures 17- 19, except for the parameter values that vary along the x-axis, all other parameters are set according to Table 6.

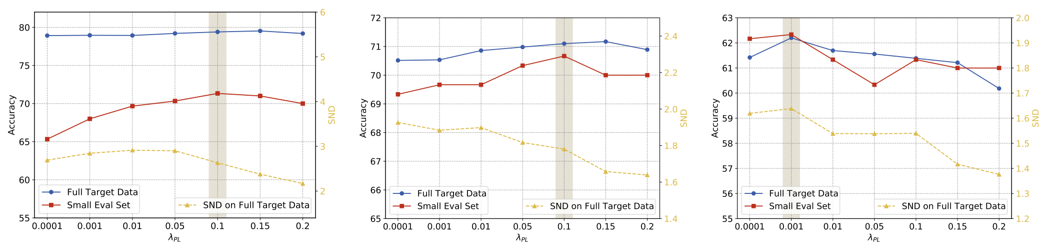
Decay Exponent β. Figure 17(a) reveals that the dispersion control term can help mitigate the sensitivity of β in contrastive learning based SFDA algorithms. Specifically, we compare the performance of an SFDA task (R to P on DomainNet-126 dataset) using our proposed method (UCon-SFDA) against the basic contrastive learning approach introduced in Yang et al. (2022). Beyond providing stable performance improvements, our method demonstrates reduced sensitivity to the hyperparameter β, benefiting from the uncertainty-controlling regularizations.

Partial Label Number K_{PL} and Uncertainty Threshold τ. Figure 17(b) and (c) illustrate the sensitivity of our method to partial label number K_{PL} and uncertainty threshold τ, respectively. By comparing the performance variations on VisDA-RUST, Office-31, and Office-Home (Pr to Cl task) under different K_{PL} and τ, we observe that the accuracy of our method is not significantly affected



(a) Decay Exponent β (b) Partial Label Num K_{PL} (c) Uncertainty Threshold τ

Figure 17: *Sensitivity analysis of the proposed method relative to different values of hyperparameters β, K_{PL}, and τ. In the legend, "wo" is the abbreviation for "without".*



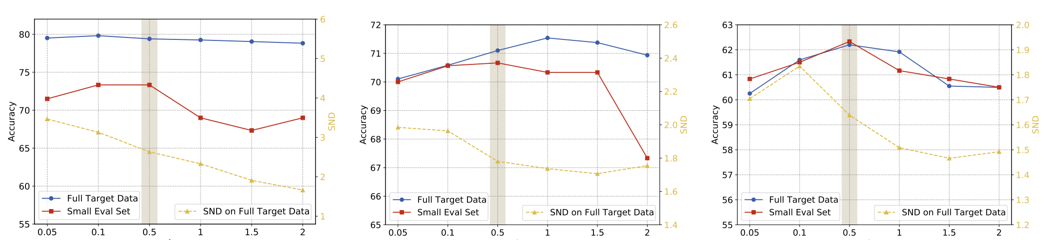
(a) VisDA-RUST (b) DomainNet-126 (R → P) (c) Office-Home (Pr → Cl)

Figure 18: *Sensitivity analysis of dispersion control loss coefficient λ_{PL}. Different colors represent various criteria for hyperparameter selection, while the shaded area indicates the parameter values chosen corresponding to the results reported in the main paper.*

by varying values of K_{PL} and τ. Moreover, the performance improvements by the partial label loss are both evident and stable (as shown by the comparison between the solid and dashed lines).

Partial Labeling term coefficient λ_{CL} and Dispersion Control term coefficient λ_{DC}. As shown in Figures 18- 19, we conducted an ablation study with finer-grained variations of λ_{CL} and λ_{DC} on three datasets to access sensitivity of the experimental results. Relative to the blue lines, the adaptation performance remains stable and robust across different values of these two hyperparameters, with the regions of optimal performance being well-concentrated.

Additional Insights for Advanced and Practical Hyperparameter Selection Strategies. Hyperparameter tuning in SFDA poses significant challenges due to the lack of target labels and substantial distribution shifts across domains. In our experiments, we found that SND scores often fail to correlate consistently with performance on the full target dataset. Moreover, sensitivity analysis based on



(a) VisDA-RUST (b) DomainNet-126 (R → P) (c) Office-Home (Pr → Cl)

Figure 19: *Sensitivity analysis of dispersion control loss coefficient λ_{DC}. Different colors represent various criteria for hyperparameter selection, while the shaded area indicates the parameter values chosen corresponding to the results reported in the main paper.*

the full target data incurs high computational costs, making it less feasible for real-world applications. To overcome these limitations, we explore a novel small evaluation set-based method. Specifically, we randomly select a subset (300 data points) from the full unlabeled target data (typically containing 5k-50k data points), manually label it, and create a pseudo-validation set. Hyperparameters are subsequently selected based on their performance on this small evaluation set. While this approach requires some manual annotation, the amount of labeled data needed is minimal, making it both practical and effective for real-world scenarios, while improving the accuracy of hyperparameter selection.

Figure 18 and Figure 19 demonstrate that the performance on the small human-labeled evaluation set (red lines) aligns more closely with the desired model performance (blue lines). In contrast, the SND score (yellow lines), which is based on feature space similarity and self-prediction entropy, sometimes fails to identify the optimal hyperparameters.

Better Performance with Finer-Grained Hyperparameter Ranges. Refining the parameter selection range (as shown Figure 18(a)-(b)) or adopting a different tuning order (e.g., tuning the partial label term first, followed by the dispersion control term, as shown in Figure 19(a)-(b)) can achieve even better results, as indicated by the highest points on the blue lines. For instance, while we initially reported the UCon-SFDA performance of 79.4 on VisDA-RUST (with $L_{\text{PL}} = 0.1$ and $L_{\text{DC}} = 0.5$), we found that using a slightly smaller $L_{\text{DC}} = 0.1$ improved its performance to 79.82. These findings demonstrate that satisfactory performance of our approach does not depend on excessive hyperparameter tuning, and further highlights the robustness and effectiveness of our algorithm.

C.6 Different Losses for Dispersion Control Term

We evaluate the performance of the dispersion control term under different similarity metrics between the anchor data point and its augmented version, $\mathbb{d}_{\boldsymbol{\theta}}\left(\texttt{AUG}\left(\mathbf{x}_i\right), \mathbf{x}_i\right)$, in Equation (7).

Specifically, for the Equation (7) in the main text, we define:

$$\mathbb{d}_{\boldsymbol{\theta}}\left(\texttt{AUG}\left(\mathbf{x}_i\right), \mathbf{x}_i\right) \triangleq \langle f_{\text{T}}(\mathbf{x}_i;\boldsymbol{\theta}), \log f_{\text{T}}\left(\texttt{AUG}\left(\mathbf{x}_i\right);\boldsymbol{\theta}\right)\rangle.$$

To further validate the role of data augmentation from the perspective of negative sampling uncertainty, we experimented with different similarity metrics, including the direct dot product and the L^2 norm, given by

$$\mathbb{d}_{\boldsymbol{\theta},\text{dot}}\left(\texttt{AUG}\left(\mathbf{x}_i\right), \mathbf{x}_i\right) \triangleq \langle f_{\text{T}}(\mathbf{x}_i;\boldsymbol{\theta}), f_{\text{T}}\left(\texttt{AUG}\left(\mathbf{x}_i\right);\boldsymbol{\theta}\right)\rangle,$$

and

$$\mathbb{d}_{\boldsymbol{\theta},\text{L}^2}\left(\texttt{AUG}\left(\mathbf{x}_i\right), \mathbf{x}_i\right) \triangleq \| f_{\text{T}}(\mathbf{x}_i;\boldsymbol{\theta}) - f_{\text{T}}\left(\texttt{AUG}\left(\mathbf{x}_i\right);\boldsymbol{\theta}\right)\|^2.$$

Additional experimental results, reported in Table 8, demonstrate the importance of treating data augmentations as negative samples as well as the effectiveness of the proposed dispersion control term. Furthermore, while the proposed $\mathbb{d}_{\boldsymbol{\theta}}$ achieves the best performance across most datasets, other loss formulations also present comparable results. These experimental observations provide guidance on effectively leveraging data augmentations in SFDA and verify the generalizability of our algorithm.

Table 8: Classification Accuracy (%) Under different Distance Measurements in Dispersion Control term. Bold text indicates the best results, and underlined text represents results that outperform the baseline.

Methods	Office-Home (Pr → Cl)	VisDA-RUST	DomainNet126 (R → P)
$\mathcal{L}_{\text{CL}}$	57.90	75.50	67.80
$\mathcal{L}_{\text{CL}} + \mathcal{L}_{\text{DC}}^{-}$ with $\mathbb{d}_{\boldsymbol{\theta}}$	<u>59.70</u>	**78.90**	**70.30**
$\mathcal{L}_{\text{CL}} + \mathcal{L}_{\text{DC}}^{-}$ with $\mathbb{d}_{\boldsymbol{\theta},\text{dot}}$	**60.21**	<u>78.02</u>	<u>70.08</u>
$\mathcal{L}_{\text{CL}} + \mathcal{L}_{\text{DC}}^{-}$ with $\mathbb{d}_{\boldsymbol{\theta},\text{L}^2}$	<u>59.14</u>	<u>77.77</u>	<u>69.34</u>

Table 9: Comparison of Training Time, Memory Usage, and Accuracy on VisDA2017.

Method	Training Time (Normalized w.r.t. AaD)	Memory Usage (Normalized w.r.t. AaD)	Accuracy (%)
AaD	1.000	1.000	87.3
SF(DA)^2	1.036	1.052	88.1
UCon-SFDA (Ours)	1.058	1.112	89.6

C.7 Training time and Resource Usage Analysis

To further validate the practical value of our proposed methodology, we conduct the training time and resource usage analysis in this subsection.

Compared to the baseline model, AaD (Yang et al., 2022), a widely utilized contrastive learning and memory bank-based SFDA method, our UCon-SFDA introduces explicit data augmentation and an additional partial label bank component. These additions increase both resource usage and computational complexity. However, such trade-offs are consistent with recent trends in the field (Hwang et al., 2024; Karim et al., 2023; Mitsuzumi et al., 2024a), where enhanced resource utilization is commonly accepted to achieve significant performance improvements.

The computational complexity of our approach remains comparable to other modern techniques that leverage data augmentation or consistency regularization. For instance, compared to Karim et al. (2023) and Mitsuzumi et al. (2024a), which also incorporate explicit data augmentation during training, our UCon-SFDA avoids relying on additional network structures. Moreover, the partial label bank only incurs a small additional memory overhead that scales linearly with the size of the target domain data, making it practical for real-world SFDA applications.

Importantly, our method demonstrates superior performance, as evidenced by the experimental results presented in the main paper. For a detailed comparison, we analyzed the training time and GPU memory usage of UCon-SFDA against AaD and SF(DA)^2 Yang et al. (2022); Hwang et al. (2024). As shown in **Table 9**, the evaluation results on VisDA2017 further validate that, with tolerable computational and storage overhead, our method achieves superior performance.