One-step Label Shift Adaptation via Robust Weight Estimation

Ruidong Fan[1], **Xiao Ouyang**[1], **Tingjin Luo**[1*], **Lijun Zhang**[2] **and Chenping Hou**[1*]
[1] National University of Defense Technology, Changsha, 410073, China
[2] Nanjing University, Nanjing, 210023, China
fanruidong1996@163.com, {ouyangxiao98, tingjinluo}@hotmail.com, zljzju@gmail.com, hcpnudt@hotmail.com

Abstract

Label shift is a prevalent phenomenon encountered in open environments, characterized by a notable discrepancy in the label distributions between the source (training) and target (test) domains, whereas the conditional distributions given the labels remain invariant. Existing label shift methods adopt a two-step strategy: initially computing the importance weight and subsequently utilizing it to calibrate the target outputs. However, this conventional strategy overlooks the intricate interplay between output adjustment and weight estimation. In this paper, we introduce a novel approach termed as One-step Label Shift Adaptation (OLSA). Our methodology jointly learns the predictive model and the corresponding weights through a bi-level optimization framework, with the objective of minimizing an upper bound on the target risk. To enhance the robustness of our proposed model, we incorporate a debiasing term into the upper-level classifier training and devise a regularization term for the lower-level weight estimation. Furthermore, we present theoretical analyses about the generalization bounds, offering guarantees for the model's performance. Extensive experimental results substantiate the efficacy of our proposal.

1 Introduction

The success of traditional machine learning methodologies is generally contingent on the closed environment hypothesis, which presumes that the training and testing data are independently and identically distributed [Mohri, 2018; Bengio *et al.*, 2021]. However, real-world learning tasks frequently occur in open environments [Zhou, 2022], where data distributions may undergo temporal variations, potentially resulting in significant performance degradation of closed-environment systems [Sugiyama and Kawanabe, 2012; Huang and Ren, 2024]. Label shift [Fan *et al.*, 2023; Li *et al.*, 2024; Fan *et al.*, 2024a], which represents a typical scenario in open environments, assumes that the label distributions in the source and target domains are distinct ($P_s(Y) \neq P_t(Y)$), whereas the conditional distributions given the labels remain consistent ($P_s(X|Y) = P_t(X|Y)$). To illustrate, in the realm of COVID-19 diagnosis, incidence rates fluctuate across diverse regions, yet the symptomatic manifestations of pneumonia remain consistent. Another pertinent instance concerns bird identification, as exemplified in Figure 1. The migratory patterns of birds can induce seasonal fluctuations in their distribution within the same geographical area. Notably, despite these seasonal shifts, their morphological characteristics remain invariant. Given the aforementioned examples, label shift demonstrates considerable potential for application, thereby emerging as a pivotal area of research in recent years [Wei *et al.*, 2024; Luo and Ren, 2024].

To mitigate the consequences of label shift and attain precise target output results, existing methodologies typically employ a two-step strategy to address the label shift issue [Zhao *et al.*, 2021; Tian *et al.*, 2023]. Initially, they calculate importance weight, which is subsequently followed by the adjustment of posterior probabilities for target samples. In the first step, certain research endeavors utilize a pre-trained source classifier to approximate the confusion matrix and class probabilities, and estimate the importance weight based on distribution transformation theory [Lipton *et al.*, 2018; Azizzadenesheli *et al.*, 2019]. Meanwhile, other studies derive the importance weight by minimizing Kullback-Leibler (KL) divergence between the weighted source distribution $P_s^w(X)$ and the target distribution $P_t(X)$ [Alexandari *et al.*, 2020; Sipka *et al.*, 2022]. Additionally, some methods aim to enhance the accuracy of weight estimation by incorporating the prior parameters of the target distribution [Sulc and Matas, 2019; Ye *et al.*, 2024b]. In the second step, existing approaches generally involve either direct or indirect adjustment of the posterior probabilities for target samples. Specifically, indirect adjustment methods leverage the equivalence between weighted source and target loss expectations to minimize the importance-weighted empirical risk, thereby obtaining the final target classifier and correcting the target outputs [Garg *et al.*, 2020; Fan *et al.*, 2024a]. On the other hand, direct adjustment methods combine the importance weight with the outputs generated by the source classifier, ensuring unbiased target outputs directly under the framework of distribution transformation theory [Wen *et al.*, 2024].

Although the aforementioned methods have demonstrated outstanding performance within the label shift framework, there are at least two critical challenges that require careful consideration. Firstly, when tackling the label shift problem,

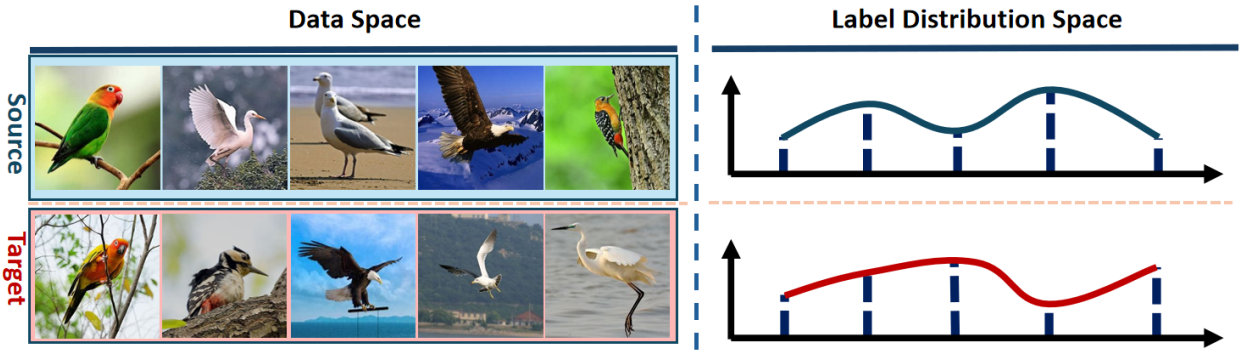


Figure 1: An illustrative instance of label shift can be observed within the domain of bird identification. During the cold winter season, both the egrets and seagulls migrate to warmer southern regions, while parrots, eagles, and woodpeckers do not engage in migration. Consequently, in the same area, the distribution of bird species varies with different seasons ($P_s(Y) \neq P_t(Y)$). However, the appearance of birds remains constant regardless of seasonal changes ($P_s(X|Y) = P_t(X|Y)$).

precise importance weight is instrumental in deriving accurate target outputs, and vice versa. However, prevalent label shift methods typically follow a sequential process of weight estimation followed by output adjustment, overlooking the intricate interplay between these two components. Secondly, numerous traditional methodologies are heavily contingent upon the efficacy of the pre-trained source classifier; in instances where this classifier exhibits suboptimal performance, the resultant bias in weight estimation can be substantial, ultimately culminating in a deterioration of the final target outputs.

In this paper, to effectively address the aforementioned challenges, we propose a novel bi-level approach denominated One-step Label Shift Adaptation (OLSA) via Robust Weight Estimation, as shown in Figure 2. Specifically, the upper-level task minimizes the unbiased target loss utilizing importance weight, whereas the lower-level objective estimates this importance weight based on target outputs. Therefore, weight estimation and classifier training can be jointly done through a bi-level optimization strategy. To enhance the robustness of our proposed model, we incorporate a bias correction term in the upper-level task. Under mild assumptions, theoretical analysis demonstrates that OLSA can attain the same expected risk as conventional label shift methods. Furthermore, we introduce a regularization term in the lower-level objective to mitigate the dependency of weight estimation on the trained classifier, thereby ensuring stable performance across diverse label shift scenarios. We derive a generalization bound for the final classifier. Finally, comprehensive experimental results are presented to validate the efficacy of OLSA. In summary, the contributions of our research are outlined as follows:

- Our research focuses on developing a method to address the label shift issue in a single step. This strategy avoids the necessity for intermediary steps, thereby enabling a more streamlined and efficient process for training the target classifier and estimating importance weight.
- To enhance the robustness of our proposed model, we incorporate a debiasing term into the upper-level task and devise a regularization term for the lower-level objective. Under mild assumptions, theoretical proof has been provided for the efficacy of our approach.
- We demonstrate the effectiveness of our approach across diverse datasets. The experimental results consistently indicate that our approach outperforms other comparative methods in the majority of cases, particularly in scenarios involving significant label shifts.

2 Related Work

Label shift is a prevalent scenario in open environments, sparking significant interest among researchers [Tasche, 2017; Bai *et al.*, 2022]. This phenomenon occurs when there exists a disparity in label distributions between the source and target domains, while preserving the same conditional distribution [Wu *et al.*, 2021; Fan *et al.*, 2023]. Previous works, such as BBSE [Lipton *et al.*, 2018] and RLLS [Azizzadenesheli *et al.*, 2019], estimate importance weight using the confusion matrix and predicted target labels, followed by retraining a new target classifier within the Weighted Empirical Risk Minimization (WERM) framework. In addition, MLLS [Alexandari *et al.*, 2020] and SCML [Sipka *et al.*, 2022] estimate importance weight by minimizing the KL divergence between the weighted source and target distributions, and directly correct the target outputs through a posterior adjustment strategy. Building on these foundations, MAP [Sulc and Matas, 2019] and MAPLS [Ye *et al.*, 2024b] introduce the Bayesian posterior of the target label distribution parameters given data. From a novel perspective of matching the source label distribution, CPM [Wen *et al.*, 2024] maintains the same theoretical guarantees as traditional feature distribution matching frameworks, while significantly enhancing computational efficiency due to the direct matching of label variables. However, the aforementioned approaches employ a two-step strategy, which neglects the interaction between output adjustment and weight estimation.

One-step Domain Adaptation is designed to directly learn an effective target classifier under distribution shift. Vapnik's principle [Vapnik, 2013] emphasizes the avoidance of solving a general problem as an intermediary step when confronted with limited information, thereby highlighting the significance of one-step approach. DAOT [Peng *et al.*, 2018] conceptualizes the feature distribution alignment process as a one-step

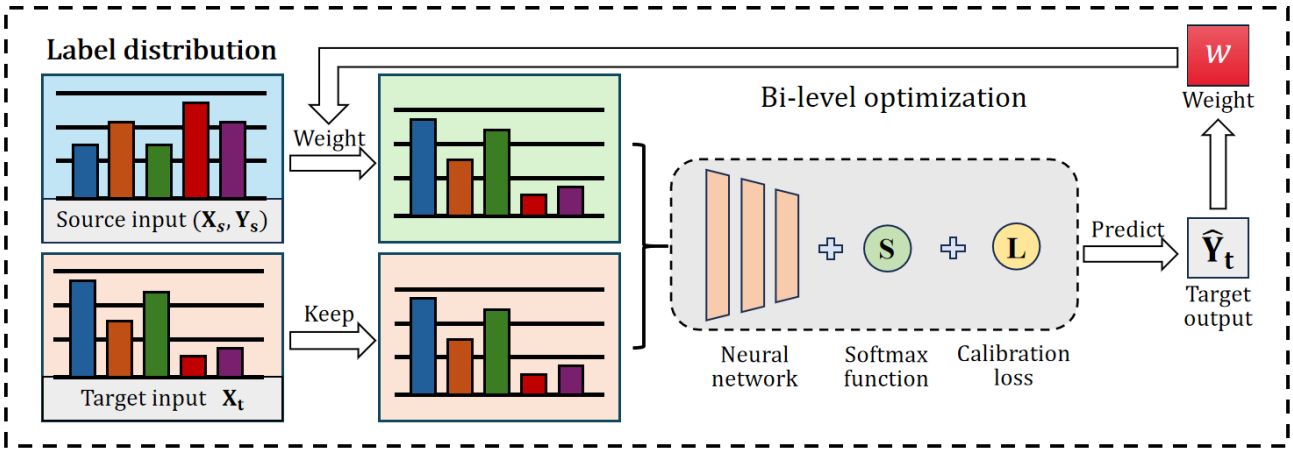


Figure 2: Illustration of the OLSA framework. We employ source inputs, target inputs, and importance weight to train the target classifier, while leveraging target outputs to estimate the importance weight. Both components are jointly optimized through a bi-level strategy, ultimately achieving a dynamic equilibrium.

transformation and implements it through a single-layer convolutional neural network. OCSA [Zhang *et al.*, 2020] introduces a one-step strategy for accommodating covariate shifts, without the intermediate step of estimating the ratio between training and test input densities. Despite these advancements, the above methods have not taken into account how to perform one-step learning in the label shift scenario. ADM-OS [Fan *et al.*, 2024b] emerges as a tailored one-step method for label shift, which jointly learns the predictive model and importance weight. However, empirical findings have indicated that ADM-OS may exhibit limited robustness under specific shifting conditions.

3 Problem Setting

3.1 Traditional WERM Framework

We employ stochastic variables $X \in \mathcal{X}, Y \in \mathcal{Y}$ to model the features and labels respectively, where $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y}$ is a discrete set as $\{1, 2, ..., K\}$. d is the feature dimension, and K represents the number of categories. We have access to labeled source data $(X_s, Y_s) = \{x_i, y_i\}_{i=1}^n$ and unlabeled target data $X_t = \{x_j\}_{j=1}^m$, which are independently and identically drawn from the source distribution P_s and target distribution P_t respectively. We define $L(\cdot, \cdot) : \Delta_{K-1} \times \mathcal{Y} \to \mathbb{R}^1$ as the loss function and $h \in \mathcal{H}$: $\mathcal{X} \mapsto \Delta_{K-1}$ as the classifier, where Δ_{K-1} denotes the standard K-dimensional probability simplex. Under the label shift assumption, i.e., $P_s(X|Y) = P_t(X|Y)$ and $P_s(Y) \neq P_t(Y)$, if we give a hypothesis space $\mathcal{H}$ and a loss function L, the goal of label shift setting is to find an optimal target classifier $h_t \in \mathcal{H}$ which minimizes the following weighted loss [Lipton *et al.*, 2018; Azizzadenesheli *et al.*, 2019; Garg *et al.*, 2020; Fan *et al.*, 2024a]:

$$\begin{aligned} \mathcal{R}_{\mathrm{T}}(h; w^*) &= \mathbb{E}_{(X,Y)\sim P_t}[L(h(X), Y)] \\ &= \mathbb{E}_{(X,Y)\sim P_s}[w^*(Y)L(h(X), Y)], \end{aligned} \tag{1}$$

where $w^* \in \mathbb{R}^K$ is the importance weight and $w^*(Y) = P_t(Y)/P_s(Y)$. In practical scenarios, the veritable weight w^* is commonly undisclosed. If w^* is approximated as $\hat{w}$, the target classifier can be derived by minimizing the following empirical loss:

$$\hat{\mathcal{R}}_{\mathrm{T}}(h; \hat{w}) = \frac{1}{n}\sum_{i=1}^{n} \hat{w}(y_i)L(h(x_i), y_i). \tag{2}$$

3.2 The OLSA Approach

In this section, we aim to address the label shift issue using a one-step strategy. Inspired by the theory of hyper-parameter optimization [Liu *et al.*, 2022; Liu *et al.*, 2024b], we treat the importance weight $\hat{w}$ as a hyper-parameter to be optimized, and then introduce a bi-level optimization technique for its solution. Specifically, the upper-level loss is employed for training the target classifier, while the lower-level loss is utilized for estimating importance weight. Now, let us introduce the specific forms of upper-level and lower-level losses respectively.

Upper-level Loss. Existing label shift methods for estimating importance weight require the availability of target outputs. Motivated by this observation, we attempt to investigate an innovative approach that incorporates unlabeled target data into training loss. The most prevalent approach is semi-supervised learning [Schmutz *et al.*, 2023; Ye *et al.*, 2024a], which leverages both labeled and unlabeled data simultaneously to train the classifier. However, due to the discrepancy between the source and target distributions, it is unreasonable to directly apply traditional methods. Therefore, we introduce the following weighted semi-supervised loss:

$$\hat{\mathcal{R}}_{\mathrm{M}}(h; \hat{w}) = \hat{\mathcal{R}}_{\mathrm{T}}(h; \hat{w}) + \frac{\beta}{m}\sum_{j=1}^{m} H_t(h(x_j)), \tag{3}$$

where $H_t(\cdot) := \int_Y P_t(Y|\cdot)L(h(\cdot), Y)\mathrm{d}Y$ represents a label-independent component, while $\beta \in (0, 1)$ is a balance parameter that governs the equilibrium between the labeled and unlabeled terms. However, the introduction of the second item naturally increases the risk of traditional label shift loss

$\hat{\mathcal{R}}_{\mathrm{T}}$. Therefore, we attempt to leverage labeled data on label-independent term to mitigate potential bias and design the following upper-level loss:

$$\hat{\mathcal{R}}_{\mathrm{DeM}}(h;\hat{w}) = \frac{1}{n}\sum_{i=1}^{n}\hat{w}(y_i)L(h(x_i),y_i) + \left(\frac{\beta}{m}\sum_{j=1}^{m}H_t(h(x_j)) - \frac{\beta}{n}\sum_{i=1}^{n}H_s(h(x_i);\hat{w})\right), \tag{4}$$

where $H_s(\cdot;\hat{w}) := \hat{w}(Y)\int_Y P_s(Y|\cdot)L(h(\cdot),Y)\mathrm{d}Y$ and the source conditional distribution $P_s(Y|\cdot)$ can be approximated by the output of source classifier with the softmax layer. Fortunately, our approach achieves the same expected risk as traditional label shift methods, as demonstrated in Theorem 1.

Theorem 1. *Assume the importance weight $\hat{w}$ reaches its optimal value $P_t(Y)/P_s(Y)$, we have*

$$\mathbb{E}\left[\hat{\mathcal{R}}_{\mathrm{DeM}}(h;\hat{w})\right] = \mathbb{E}\left[\hat{\mathcal{R}}_{\mathrm{T}}(h;\hat{w})\right]. \tag{5}$$

The comprehensive proof of the aforementioned theorem is provided in supplementary materials. On this basis, given the assumption that the importance weight $\hat{w}$ is known, we can directly optimize the loss function $\hat{\mathcal{R}}_{\mathrm{DeM}}(h;\hat{w})$ in order to obtain the target classifier. Subsequently, the critical task is to determine an accurate approach for estimating the importance weight $\hat{w}$.

Lower-level Loss. In the first, according to the properties of distributions, we give a simple estimation strategy for the weights $\hat{w}$:

$$\hat{w} = \frac{P_t(Y)}{P_s(Y)} = \frac{\sum_{j=1}^{m}P_t(Y|x_j)}{mP_s(Y)}. \tag{6}$$

Since the labels of the source data are known, we can estimate the source label distribution $P_s(Y)$ through statistical class frequency, i.e.,

$$P_s(Y=k) = \frac{1}{n}\sum\nolimits_{i=1}^{n}\mathbf{1}\{Y_i=k\}, k\in[K]. \tag{7}$$

On this basis, the key issue becomes how to estimate the conditional distribution $P_t(Y|X)$. A classic approach is to approximate the true conditional distribution by considering the output of softmax layer in a deep neural network, i.e., $h_t(x)\approx P_t(Y|X)$. However, previous studies [Guo *et al.*, 2017; Liu *et al.*, 2024a] have demonstrated that conventional neural networks encounter challenges in accurately estimating class posterior probabilities, despite their proficiency as classifiers. Most label shift works use bias-corrected temperature scaling (BCTS) [Alexandari *et al.*, 2020; Sipka *et al.*, 2022] to calibrate outputs. However, as it is a post-processing technique, it is not applicable to our one-step approach. Therefore, for good classifier calibration, we introduce γ-loss, which is define as follows:

$$L\left(h(x),y;\gamma\right) = \begin{cases} -\log\left(h_y(x)\right), & \gamma=1; \\ \frac{\gamma}{\gamma-1}\left(1-h_y(x)^{1-\frac{1}{\gamma}}\right), & \text{others.}\end{cases} \tag{8}$$

By minimizing the expectation of the above loss [Sypherd *et al.*, 2022], we derive the optimal outputs as

$$h_k^*(x) = \frac{P(k|x)^\gamma}{\sum_i P(i|x)^\gamma}, \quad \forall k\in[K]. \tag{9}$$

The parameter γ has the ability to 'soften' the outputs. When $\gamma=1$, the output is consistent with traditional softmax output, and as $\gamma\to 0$, the output $h(x)$ gravitates towards $1/K$, indicating maximum uncertainty. By combining Eq. (6), we estimate the importance weight $\hat{w}$ as

$$\hat{w} = \frac{\sum_{j=1}^{m}P_t(Y|x_j)}{mP_s(Y)} = \frac{\sum_{j=1}^{m}h(x_j)}{mP_s(Y)} \triangleq T(h). \tag{10}$$

Thus, we can design the lower-level loss as follows:

$$\hat{w} = \arg\min\nolimits_{w\succeq 0, w^{\mathrm{T}}P_s(Y)=1}\|w-T(h)\|_2^2. \tag{11}$$

However, the aforementioned estimation is heavily influenced by the classifier outputs $h(x)$. If the calibration effect is poor, it can lead to significant weight bias. Thus, we try to add a regularization term to alleviate this situation.

$$\hat{w} = \underset{w\succeq 0, w^{\mathrm{T}}P_s(Y)=1}{\arg\min}\|w-T(h)\|_2^2 + \lambda\left\|w-\frac{1}{KP_s(Y)}\right\|_2^2. \tag{12}$$

Where λ controls how much the label distribution is skewed. When $\lambda\to\infty$, $w=\frac{1}{KP_s(Y)}$ and $P_t(Y)=\frac{1}{K}$. This regularization term ensures that the target distribution is balanced even when the estimation deviation is large, thus ensuring the stability of training. Through the transformation of Eq. (12), we get the final lower-level loss as

$$\hat{w} = \arg\min\nolimits_{w\succeq 0, w^{\mathrm{T}}P_s(Y)=1}\left\|w-\frac{T(h)+\lambda\frac{1}{KP_s(Y)}}{1+\lambda}\right\|_2^2. \tag{13}$$

Total Loss. By combining the upper-level loss Eq. (4) and lower-level loss Eq. (13), we have the total loss as follows:

$$\begin{cases} \min\limits_{h}\hat{\mathcal{R}}_{\mathrm{DeM}}(h,\hat{w}) = \frac{1}{n}\sum\limits_{i=1}^{n}\hat{w}(y_i)L(h(x_i),y_i;\gamma) \\ \quad + \left(\frac{\beta}{m}\sum\limits_{j=1}^{m}H_t(h(x_j);\gamma) - \frac{\beta}{n}\sum\limits_{i=1}^{n}H_s(h(x_i);\gamma,\hat{w})\right), \\ s.t.\ \hat{w} = \arg\min_{w\succeq 0, w^{\mathrm{T}}P_s(Y)=1}\left\|w-\frac{T(h)+\lambda\frac{1}{KP_s(Y)}}{1+\lambda}\right\|_2^2. \end{cases} \tag{14}$$

Then, we prove an upper bound for the generalization error of OLSA as follows.

Theorem 2. *Give n samples drawn from the source distribution P_s, m samples drawn from the target distribution P_t and bound loss function L and H. Then, there exists two constants $\kappa_1,\kappa_2>0$, that depends on β, w^*, L and H, the following generalization bound holds with probability at least $1-\delta$,*

$$\mathcal{R}_{\mathrm{T}}(h;w^*) \le \hat{\mathcal{R}}_{\mathrm{DeM}}(h;\hat{w}) + 2\mathrm{Rad}_{n+m} + \kappa_1||\hat{w}-w^*||_2 + \kappa_2\sqrt{\frac{\log(4/\delta)}{n+m}}, \tag{15}$$

where Rad_{n+m} is the Rademacher complexity.

The detailed proof can be found in supplementary materials.

3.3 Optimization

The aforementioned loss constitutes a bi-level problem, necessitating the utilization of an optimization framework grounded in implicit gradients (IG). This strategy enables precise modeling and optimization of the intricate interplay between the estimated weight $\hat{w}$ and model parameters h. Specifically, the IG derived from Eq. (15) can be mathematically formulated as follows:

$$\nabla\hat{\mathcal{R}}_{\mathrm{DeM}}(h,\hat{w}) = \nabla_h\hat{\mathcal{R}}_{\mathrm{DeM}}(h,\hat{w}) + \underbrace{\frac{\mathrm{d}\hat{w}^{\mathrm{T}}}{\mathrm{d}h}}_{\mathrm{IG}}\nabla_{\hat{w}}\hat{\mathcal{R}}_{\mathrm{DeM}}(h,\hat{w}), \tag{16}$$

where ∇_h and $\nabla_{\hat{w}}$ represent the partial derivatives of the bivariate function. Traditional bi-level optimization approaches derive the IG formula through the meticulous application of implicit function theory. However, the computation of IG presents inherent difficulties, primarily arising from the complexities involved in matrix inversions, second-order partial derivatives, and constraints. In order to effectively compute IG, it is evident that the optimal solution $\hat{w}^*$ of the unconstrained lower-level loss function is formulated as follows:

$$\hat{w}^* = \frac{\frac{\sum_{j=1}^{m} h(x_j)}{mP_s(Y)} + \lambda\frac{1}{KP_s(Y)}}{1+\lambda}. \tag{17}$$

Since we adopt the softmax layer in our model, the output $h(x)$ satisfies that $h(x) \succeq 0$ and $\sum_{i=1}^{K} h_i(x) = 1$ for any $x \sim P_t(X)$. Thus we have

$$\begin{cases} \hat{w}^* \succeq 0, \\ \hat{w}^{*\mathrm{T}} P_s(Y) = \sum_{i=1}^{K} \frac{\frac{\sum_{j=1}^{m} h_i(x_j)}{m} + \lambda\frac{1}{K}}{1+\lambda} = 1. \end{cases} \tag{18}$$

On this basis, we can derive the optimal solution for the constrained lower-level loss as $\hat{w}^*$, and compute the IG ($\mathrm{d}\hat{w}^{*\mathrm{T}}/\mathrm{d}h$) directly. In conclusion, the procedure for optimizing Eq. (14) is summarized in Algorithm 1.

Algorithm 1 Procedure of OLSA approach

Input: The labeled source samples $\{x_i, y_i\}_{i=1}^{n}$, unlabeled target samples $\{x_i\}_{i=n+1}^{n+m}$, trade-off parameters λ, β and γ, upper-level learning rate υ and a pre-trained source classifier $h_s(x) \approx P_s(Y|x)$.
Initialize: Initialize network parameters $h_0 = h_s$;
for Iteration $j = 1, ..., T$ do
  Lower-level: Update the weight $\hat{w}_j$ by optimizing Eq. (17);
  Upper-level: Update the network parameters $\hat{h}_j$ via stochastic gradient descent calling Eq. (16);
end for
  Use the updated classifier $\hat{h}_T$ to make predictions $\{\hat{y}_i\}_{i=n+1}^{n+m}$ on the unlabeled target samples;
Output: The ultimate classifier $\hat{h}_T$ and the predicted labels $\{\hat{y}_i\}_{i=n+1}^{n+m}$.
End procedure

4 Experiments

In this section, we commence by introducing the datasets, comparative methodologies, network architectures, and parameter settings. Building upon these foundational elements, we undertake a comprehensive evaluation of the performance and effectiveness of the proposed OLSA approach, focusing on two key aspects. For the initial component, we conduct a comparative analysis of OLSA with traditional label shift methods across diverse shift scenarios and evaluation metrics. In the second component, we offer detailed outcomes derived from in-depth examinations of OLSA across various dimensions, encompassing an analysis of the deviation in importance weight estimation, an exploration of various parameter configurations and a visualization of label distribution bias.

4.1 Configuration

Dataset. In our study, we assess the performance and efficacy of OLSA on the MNIST [LeCun *et al.*, 1998], Fasion-MNIST [Xiao *et al.*, 2017], CIFAR10 [Krizhevsky *et al.*, 2009] and CIFAR100 [Krizhevsky *et al.*, 2009] datasets, incorporating numerous artificial shifts. Specifically, we introduce two distinct shift categories in our experimental setup: (1) Tweak-One shift, where the probability of a specific source class is altered to ρ (larger values of ρ result in more extreme label shift), while the probabilities of remaining classes maintain their original proportions. (2) Dirichlet shift, which generates a Dirichlet distribution utilizing the concentration parameter α (smaller values of α result in more extreme label shift), and aligns the source label distribution with this Dirichlet distribution. For the MNIST and Fasion-MNIST datasets, we allocate 2000 samples each for the training and validation sets, and 10,000 samples for the test set. Analogously, for the CIFAR10 dataset, we assign 4000 samples each for training and validation, and 10,000 samples for testing. For the CIFAR100 dataset, the distribution is 10,000 samples for training, 5000 for validation, and 20,000 for testing. Notably, to guarantee the robustness of training process, each class within the shift set is represented by a minimum of 40 samples.

Compared methods. In the main experiment subsection, we demonstrate the performance of OLSA by conducting a comparative analysis with the following traditional label shift techniques. (1) Wout-W illustrates the performance of the base classifier in the scenario where the estimated weight is excluded from consideration. (2) BBSE [Lipton *et al.*, 2018] and RLLS [Azizzadenesheli *et al.*, 2019] represent two label shift methods that are grounded in the utilization of hard confusion matrices, while BBSE-S and RLLS-S are based on the employment of soft confusion matrices. (3) MLLS [Alexandari *et al.*, 2020], CML [Sipka *et al.*, 2022], SCML [Sipka *et al.*, 2022], MAPLS [Ye *et al.*, 2024b] and CPMCN [Wen *et al.*, 2024] are five advanced label shift methods that incorporate classifier calibration and obtain the target outputs directly. (4) ADM-OS is a one-step label shift method that alternately iterates between importance weight estimation and classifier training.

Network architectures and evaluation indicators. We can adopt any model as the base model for the source classifier, and the choice of the model has implications for the accuracy of subsequent importance weight estimation. In our primary

Dataset	Shift Types	Wout-W	BBSE	BBSE-S	RLLS	RLLS-S	MLLS	CML	SCML	MAPLS	CPMCN	ADM-OS	OLSA
MNIST	$\alpha = 0.1$	0.8054	0.8363	0.8382	0.8364	0.8387	0.8514	0.8414	0.8425	0.8509	0.8515	0.8526	**0.8702**
	$\alpha = 0.5$	0.8468	0.8707	0.8724	0.8707	0.8724	0.8767	0.8772	0.8773	0.8731	0.8761	0.8768	**0.8878**
	$\alpha = 1.0$	0.8504	0.8796	0.8812	0.8796	0.8813	0.8860	0.8839	0.8854	0.8859	0.8861	0.8861	**0.9014**
	$\alpha = 5.0$	0.8899	0.8941	0.8943	0.8941	0.8939	0.8934	0.8935	0.8935	0.8936	0.8933	0.8993	**0.9024**
	$\rho = 0.3$	0.8971	0.9019	0.9019	0.9019	0.9019	0.9013	0.9011	0.9009	0.9013	0.9009	0.9033	**0.9079**
	$\rho = 0.5$	0.8803	0.8895	0.8900	0.8895	0.8901	0.8911	0.8908	0.8907	0.8911	0.8898	0.8931	**0.8991**
	$\rho = 0.7$	0.8559	0.8749	0.8758	0.8749	0.8757	0.8738	0.8754	0.8752	0.8747	0.8738	0.8811	**0.8860**
	$\rho = 0.9$	0.7381	0.7981	0.7993	0.7981	0.7996	0.8082	0.7976	0.7997	0.8062	0.8106	0.8155	**0.8391**
Fasion-MNIST	$\alpha = 0.1$	0.6493	0.7495	0.6892	0.7495	0.6938	0.6774	0.7226	0.7271	0.6828	0.6773	0.7590	**0.7770**
	$\alpha = 0.5$	0.6701	0.7498	0.7041	0.7509	0.7370	0.6910	0.7352	0.7396	0.7049	0.7211	0.7337	**0.7842**
	$\alpha = 1.0$	0.7383	0.7979	0.7765	0.7979	0.7969	0.7773	0.7884	0.7874	0.7762	0.7772	0.7812	**0.8084**
	$\alpha = 5.0$	0.8069	0.8195	0.8188	0.8195	0.8199	0.8084	0.8108	0.8108	0.8128	0.8082	**0.8202**	0.8187
	$\rho = 0.3$	0.7603	0.8222	0.7561	0.8222	0.8200	0.7694	0.7996	0.7977	0.7711	0.7690	0.8129	**0.8265**
	$\rho = 0.5$	0.7389	0.7805	0.7002	0.7805	0.7744	0.7677	0.7604	0.7508	0.7506	0.7681	0.7862	**0.7960**
	$\rho = 0.7$	0.6851	0.7668	0.6864	0.7668	0.7646	0.7231	0.6915	0.6693	0.7108	0.7231	0.7770	**0.7909**
	$\rho = 0.9$	0.4963	0.6299	0.5937	0.6586	0.5955	0.5653	0.5786	0.5905	0.5432	0.5962	0.7034	**0.7415**
CIFAR10	$\alpha = 0.1$	0.2156	0.3342	0.2745	0.3945	0.3285	0.2472	0.3193	0.3233	0.2517	0.2323	0.3928	**0.4581**
	$\alpha = 0.5$	0.4624	0.5548	0.5002	0.5409	0.5355	0.5248	0.5566	0.5207	0.5310	0.5247	0.5794	**0.6127**
	$\alpha = 1.0$	0.5260	0.5753	0.5644	0.5562	0.5553	0.5488	0.5882	0.5623	0.5521	0.5673	0.6100	**0.6219**
	$\alpha = 5.0$	0.5534	0.6162	0.5877	0.6299	0.6482	0.5973	0.5841	0.5866	0.5871	0.5969	0.6643	**0.6705**
	$\rho = 0.3$	0.5078	0.5278	0.5179	**0.5347**	0.5031	0.5320	0.5119	0.5187	0.5083	0.5190	0.5217	0.5252
	$\rho = 0.5$	0.4614	0.5050	0.4929	0.5060	0.4978	0.4942	0.4805	0.4904	0.5154	0.5029	0.5141	**0.5191**
	$\rho = 0.7$	0.4004	0.4724	0.3850	0.4767	0.3919	0.4557	0.4504	0.4331	0.4602	0.4554	0.4695	**0.5099**
	$\rho = 0.9$	0.3753	0.3898	0.3443	0.4135	0.3832	0.4077	0.4138	0.3970	0.4055	0.4077	0.3994	**0.4358**
CIFAR100	$\alpha = 0.1$	0.1014	0.1265	0.0704	0.1208	0.1141	0.1281	0.1260	0.1265	0.1151	0.1016	0.1221	**0.1675**
	$\alpha = 0.5$	0.1401	0.1711	0.1268	0.1786	0.1676	0.1664	0.1682	0.1735	0.1513	0.1629	0.1678	**0.2135**
	$\alpha = 1.0$	0.1846	0.2202	0.1318	0.2357	0.2080	0.2067	0.2177	0.2199	0.2147	0.2073	0.2069	**0.2727**
	$\alpha = 5.0$	0.2335	0.2939	0.2101	0.2957	0.2797	0.2593	0.2765	0.2810	0.2731	0.2689	0.2911	**0.3416**
	$\rho = 0.3$	0.2312	0.3175	0.1904	0.3336	0.2523	0.3011	0.3175	0.3092	0.3299	0.2932	0.3021	**0.3718**
	$\rho = 0.5$	0.1136	0.1884	0.1016	0.1985	0.1740	0.1779	0.1897	0.1956	0.1979	0.1735	0.1215	**0.2480**
	$\rho = 0.7$	0.0877	0.0818	0.0983	0.1378	0.1065	0.1074	0.1184	0.1229	0.0954	0.0853	0.0634	**0.1931**
	$\rho = 0.9$	0.0693	0.0683	0.0906	0.1244	0.1133	0.0877	0.1124	0.1143	0.0807	0.0681	0.0526	**0.1616**

Table 1: F-score performance (mean) comparison on Dirichlet and Tweak-One shift datasets.

experimental configuration, we employ a two-layer fully connected neural network for the MNIST and Fasion-MNIST datasets, while concurrently utilizing ResNet-18 [He *et al.*, 2016] for both the CIFAR-10 and CIFAR-100 datasets. We sample the data 5 times using the specified shift parameter and calculate the average as final outputs. To assess the effectiveness of OLSA, we employed accuracy (Acc), F-score, and mean squared error (MSE) as evaluation metrics, where MSE is defined as follows:

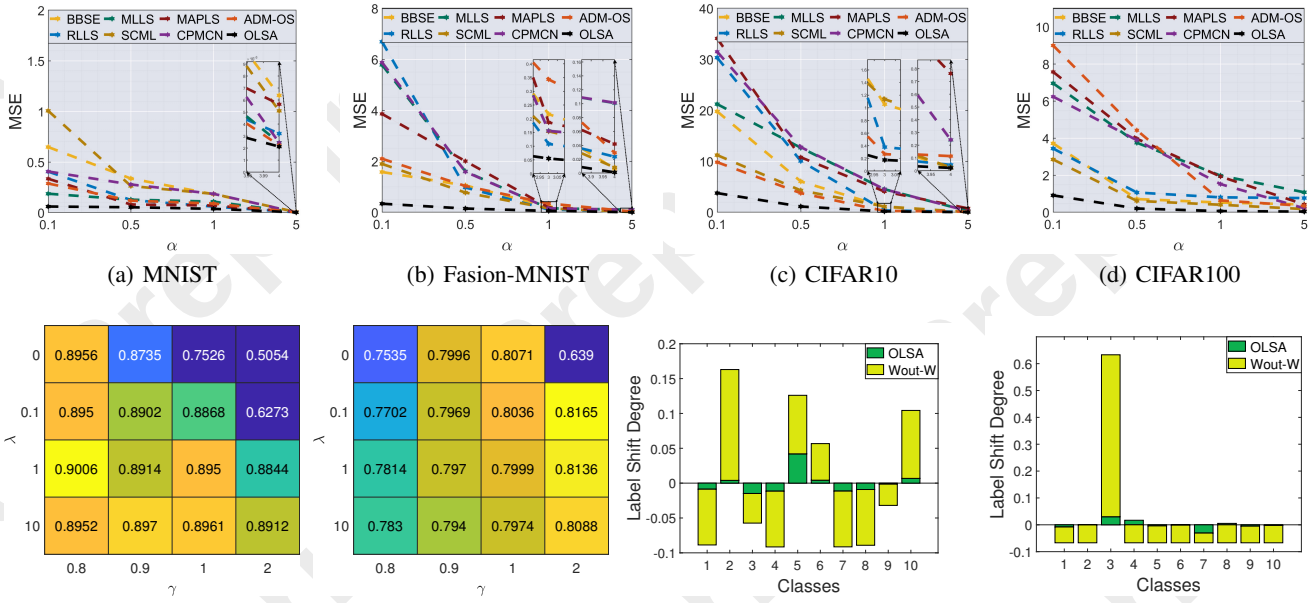
$$\mathrm{MSE}(\hat{w}) = \frac{1}{\mathrm{len}(\hat{w})} \left\| \hat{w} - \frac{P_t(Y)}{P_s(Y)} \right\|^2. \tag{19}$$

Parameter settings. The parameters for compared methods are chosen based on the technologies delineated in their respective references. In our study, we report the results of different shift parameters $\alpha \in [0.1, 0.5, 1.0, 5.0]$ and $\rho = [0.3, 0.5, 0.7, 0.9]$. In addition, we fix the trade-off parameter $\beta = 0.1$ empirically, while the calibration parameter γ is selected from the discrete set $[0.8, 0.9, 1, 2]$ and the regularization parameter λ is chosen form the discrete set $[0, 0.1, 1, 10]$ through the validation set results.

4.2 Main Results

To demonstrate the efficacy of our OLSA approach, all methods run on framework with Python 3.7 and PyTorch based on the same pre-trained classifier. We present the Acc (shown in supplementary materials) and F-score (shown in Table 1) on both the Dirichlet and Tweak-One shift datasets, and have the following observations.

1. While the performance of different comparative methods varies across diverse datasets, our OLSA approach demonstrates a consistent improvement in the performance of existing label shift methods in most cases. For example, on the Dirichlet shift CIFAR10 dataset with $\alpha = 0.1$, OLSA exhibits significant improvements, achieving an increase of nearly 6% in the F-score.
2. When compared to the performance outcomes observed under conditions of small label shifts, OLSA demonstrates a more significant enhancement in performance under scenarios involving large label shifts. For example, on the Tweak-One shift Fasion-MNIST dataset with small shift $\rho = 0.3$, OLSA exhibits a decrement of 1% in the F-score, while under large shift $\rho = 0.9$, achieves a notable increase of approximately 4% in the F-score.



(a) MNIST (b) Fasion-MNIST (c) CIFAR10 (d) CIFAR100

(e) MNIST ($\rho = 0.5$) (f) Fasion-MNIST ($\alpha = 1$) (g) MNIST ($\alpha = 0.5$) (h) CIFAR10 ($\rho = 0.7$)

Figure 3: Performance analysis diagram of OLSA.

These findings underscore the robustness and stability of OLSA in addressing large label shift scenarios.

3. It is noteworthy that, on the CIFAR100 dataset, OLSA attains a substantial improvement in the F-score, albeit with less satisfactory results in terms of Acc. This disparity arises from the significant weight estimation bias exhibited by the compared methods, which results in highly precise predictions for classes with abundant samples while overlooking the less represented classes.

4.3 Performance Analyses

MSE comparison

To demonstrate the efficacy of our shift strategy, we introduce MSE indicator, which represents the deviation between the estimated and true weights. The results are depicted in Fig. 3 (a-d), from which several key observations have been derived. (1) The performance of OLSA remains superior to compared methods in the majority of cases, especially in cases where the degree of shift is significant. This fully demonstrates the stability of our method, indicating its applicability to most shift scenarios. (2) We observe that accurate weights do not necessarily lead to a better classifier among the comparative methods. However, such a scenario rarely occurs in our approach. This indirectly demonstrates the effectiveness of OLSA, which simultaneously trains the classifier and estimates the importance weight.

Parameter sensitivity analysis

Here, we check the Acc of the trade-off parameters λ and γ on the Tweak-One shift MNIST and Dirichlet shift Fasion-MNIST datasets respectively. The results are visually presented in Fig. 3 (e-f). We find that for different datasets and shift scenarios, the optimal value of γ varies, which underscores the necessity of adjusting it, showcasing the effectiveness of the calibration loss. Furthermore, when the value of λ is large, the model's performance typically remains good, which illustrates the effectiveness of introducing the lower-level regularization term.

Visualization of label distribution bias

In order to demonstrate that OLSA approach can alleviate label shift, we present some visualization results in Figure 3 (g-h), which shows the label shift degree $\hat{P}_t(Y) - P_t(Y)$ of the Wout-W and OLSA methods. A smaller degree of label shift indicates a more accurate estimation of the target label distribution. As can be observed from figures, OLSA approach demonstrates robust performance, even under conditions of significant label shift, with a deviation from the true label distribution remaining within 0.05.

5 Conclusion

In this paper, we introduce an innovative one-step methodology to address the label shift problem, which is crucial but rarely studied. Utilizing bilevel optimization techniques, OLSA strengthens the interplay between weight estimation and classifier training. Dependent on the introduced regularization terms, OLSA approach opens up an interesting frontier for the robust one-step modeling of scenarios involving large label shifts. The solid theoretical analysis and enriched experimental analysis fully demonstrate the effectiveness of our OLSA approach. In future work, attempts can be made to enhance the robustness of weight estimation by refining the regularization term. This may be achieved by leveraging data structures, among other approaches.

Acknowledgments

This work was partially supported by the National Key Research and Development Program (No. 2022ZD0114803), the NSF for Distinguished Young Scholars under Grant No. 62425607 and the Key NSF of China under Grant No. 62136005. Chenping Hou and Tingjin Luo are the corresponding authors.

References

[Alexandari *et al.*, 2020] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *Proceedings of the 37th International Conference on Machine Learning(ICML)*, pages 222–232, 2020.

[Azizzadenesheli *et al.*, 2019] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *The 7th International Conference on Learning Representations(ICLR)*, 2019.

[Bai *et al.*, 2022] Yong Bai, Yu-Jie Zhang, Peng Zhao, Masashi Sugiyama, and Zhi-Hua Zhou. Adapting to online label shift with provable guarantees. In *Advances in Neural Information Processing Systems 35(NeurIPS)*, volume 35, pages 29960–29974, 2022.

[Bengio *et al.*, 2021] Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. Deep learning for ai. *Communications of the ACM*, 64(7):58–65, 2021.

[Fan *et al.*, 2023] Ruidong Fan, Xiao Ouyang, Tingjin Luo, Dewen Hu, and Chenping Hou. Incomplete multi-view learning under label shift. *IEEE Transactions on Image Processing*, 32:3702–3716, 2023.

[Fan *et al.*, 2024a] Ruidong Fan, Xiao Ouyang, Hong Tao, and Chenping Hou. Label shift correction via bidirectional marginal distribution matching. In *Proceedings of the 30th ACM Conference on Knowledge Discovery and Data Mining(KDD)*, pages 735–746, 2024.

[Fan *et al.*, 2024b] Ruidong Fan, Xiao Ouyang, Hong Tao, Yuhua Qian, and Chenping Hou. Theory-inspired label shift adaptation via aligned distribution mixture. *CoRR*, abs/2411.02047, 2024.

[Garg *et al.*, 2020] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary C. Lipton. A unified view of label shift estimation. In *Advances in Neural Information Processing Systems 33(NeurIPS)*, 2020.

[Guo *et al.*, 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning(ICML)*, volume 70, pages 1321–1330, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 770–778, 2016.

[Huang and Ren, 2024] Zixian Huang and Chuan-Xian Ren. Rethinking correlation learning via label prior for open set domain adaptation. In *Proceedings of the 33th International Joint Conference on Artificial Intelligence(IJCAI)*, pages 884–892, 2024.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Li *et al.*, 2024] Juren Li, Yang Yang, Youmin Chen, Jianfeng Zhang, Zeyu Lai, and Lujia Pan. DWLR: domain adaptation under label shift for wearable sensor. In *Proceedings of the 33th International Joint Conference on Artificial Intelligence(IJCAI)*, pages 4425–4433, 2024.

[Lipton *et al.*, 2018] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 3122–3130, 2018.

[Liu *et al.*, 2022] Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10045–10067, 2022.

[Liu *et al.*, 2024a] Jiawei Liu, Changkun Ye, Ruikai Cui, and Nick Barnes. Self-calibrating vicinal risk minimisation for model calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 3335–3345, 2024.

[Liu *et al.*, 2024b] Risheng Liu, Zhu Liu, Wei Yao, Shangzhi Zeng, and Jin Zhang. Moreau envelope for nonconvex bi-level optimization: A single-loop and hessian-free solution strategy. In *Proceedings of the 41th International Conference on Machine Learning(ICML)*, 2024.

[Luo and Ren, 2024] You-Wei Luo and Chuan-Xian Ren. When invariant representation learning meets label shift: Insufficiency and theoretical insights. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9407–9422, 2024.

[Mohri, 2018] Mehryar Mohri. *Foundations of machine learning*. MIT press, 2018.

[Peng *et al.*, 2018] Xishuai Peng, Yuanxiang Li, Yi Lu Murphey, Xian Wei, and Jianhua Luo. Domain adaptation with one-step transformation. In *IEEE Symposium Series on Computational Intelligence(SSCI)*, pages 539–546, 2018.

[Schmutz *et al.*, 2023] Hugo Schmutz, Olivier Humbert, and Pierre-Alexandre Mattei. Don't fear the unlabelled: safe semi-supervised learning via debiasing. In *The 11th International Conference on Learning Representations(ICLR)*, 2023.

[Sipka *et al.*, 2022] Tomás Sipka, Milan Sulc, and Jirí Matas. The hitchhiker's guide to prior-shift adaptation. In

IEEE/CVF Winter Conference on Applications of Computer Vision(WACV), pages 2031–2039, 2022.

[Sugiyama and Kawanabe, 2012] Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

[Sulc and Matas, 2019] Milan Sulc and Jiri Matas. Improving CNN classifiers by estimating test-time priors. In *IEEE/CVF International Conference on Computer Vision Workshops(ICCV Workshops)*, pages 3220–3226, 2019.

[Sypherd *et al.*, 2022] Tyler Sypherd, Mario Díaz, John Kevin Cava, Gautam Dasarathy, Peter Kairouz, and Lalitha Sankar. A tunable loss function for robust classification: Calibration, landscape, and generalization. *IEEE Transactions on Information Theory*, 68(9):6021–6051, 2022.

[Tasche, 2017] Dirk Tasche. Fisher consistency for prior probability shift. *Journal of Machine Learning Research*, 18(95):1–32, 2017.

[Tian *et al.*, 2023] Qinglong Tian, Xin Zhang, and Jiwei Zhao. ELSA: efficient label shift adaptation through the lens of semiparametric models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 34120–34142, 2023.

[Vapnik, 2013] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[Wei *et al.*, 2024] Tong Wei, Zhen Mao, Zi-Hao Zhou, Yuanyu Wan, and Min-Ling Zhang. Learning label shift correction for test-agnostic long-tailed recognition. In *Proceedings of the 41th International Conference on Machine Learning(ICML)*, 2024.

[Wen *et al.*, 2024] Hongwei Wen, Annika Betken, and Hanyuan Hang. Class probability matching with calibrated networks for label shift adaption. In *The 12th International Conference on Learning Representations(ICLR)*, 2024.

[Wu *et al.*, 2021] Ruihan Wu, Chuan Guo, Yi Su, and Kilian Q Weinberger. Online adaptation to label distribution shift. volume 34, pages 11340–11351, 2021.

[Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[Ye *et al.*, 2024a] Bo Ye, Kai Gan, Tong Wei, and Min-Ling Zhang. Bridging the gap: Learning pace synchronization for open-world semi-supervised learning. In *Proceedings of the 33th International Joint Conference on Artificial Intelligence(IJCAI)*, pages 5362–5370, 2024.

[Ye *et al.*, 2024b] Changkun Ye, Russell Tsuchida, Lars Petersson, and Nick Barnes. Label shift estimation for class-imbalance problem: A bayesian approach. In *IEEE/CVF Winter Conference on Applications of Computer Vision(WACV)*, pages 1062–1071, 2024.

[Zhang *et al.*, 2020] Tianyi Zhang, Ikko Yamane, Nan Lu, and Masashi Sugiyama. A one-step approach to covariate shift adaptation. In *Asian Conference on Machine Learning(ACML)*, pages 65–80, 2020.

[Zhao *et al.*, 2021] Eric Zhao, Anqi Liu, Animashree Anandkumar, and Yisong Yue. Active learning under label shift. In *The 24th International Conference on Artificial Intelligence and Statistics(AISTATS)*, volume 130, pages 3412–3420, 2021.

[Zhou, 2022] Zhi-Hua Zhou. Open-environment machine learning. *National Science Review*, 9(8):nwac123, 2022.