IW-GAE: Importance Weighted Group Accuracy Estimation for Improved Calibration and Model Selection in Unsupervised Domain Adaptation

Taejong Joo [1] **Diego Klabjan** [1]

Abstract

Distribution shifts pose significant challenges for model calibration and model selection tasks in the unsupervised domain adaptation problem—a scenario where the goal is to perform well in a distribution shifted domain without labels. In this work, we tackle difficulties coming from distribution shifts by developing a novel importance weighted group accuracy estimator. Specifically, we present a new perspective of addressing the model calibration and model selection tasks by estimating the group accuracy. Then, we formulate an optimization problem for finding an importance weight that leads to an accurate group accuracy estimation with theoretical analyses. Our extensive experiments show that our approach improves state-of-the-art performances by 22% in the model calibration task and 14% in the model selection task.

1. Introduction

In this work, we consider a classification problem in unsupervised domain adaptation (UDA). UDA aims to transfer knowledge from a source domain with ample labeled data to enhance the performance in a target domain where labeled data is unavailable. In UDA, the source and target domains have different data generating distributions, so the core challenge is to transfer knowledge contained in the labeled dataset in the source domain to the target domain under the distribution shifts. Over the decades, significant improvements in the transferability of accuracy from source to target domains have been made, resulting in areas like domain alignment (Ben-David et al., 2010; Zhang et al., 2019) and self-training (Chen et al., 2020; Cai et al., 2021).

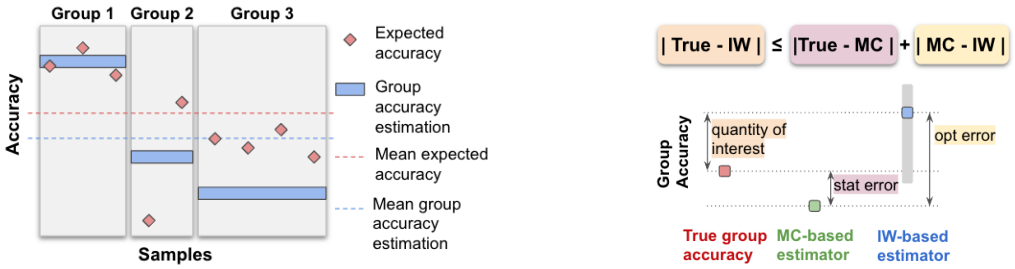
[1]Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL, USA. Correspondence to: Taejong Joo <taejong.joo@northwestern.edu>, Diego Klabjan <d-klabjan@northwestern.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

However, model calibration, which is about matching prediction confidence on a sample to its expected accuracy (Dawid, 1982; Guo et al., 2017), remains challenging in UDA due to the distribution shifts. Specifically, it is widely known that state-of-the-art calibrated classifiers in the independent and identically distributed (i.i.d.) settings (Guo et al., 2017; Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017) begin to generate over-confident predictions in the face of distributional shifts (Ovadia et al., 2019). Further, Wang et al. (2020) show the discernible compromise in calibration performance as an offset against the enhancement of the accuracy in the target domain.

Moreover, the model selection task in UDA remains challenging due to the scarcity of labeled target domain data that are required to evaluate model performance. In the i.i.d. settings, a standard approach for model selection is a cross-validation method—constructing a hold-out dataset for selecting the model that yields the best performance on the hold-out dataset. While cross-validation provides favorable statistical guarantees (Stone, 1977; Kohavi et al., 1995), such guarantees falter in the presence of the distribution shifts due to the violation of the i.i.d. assumption. In practice, it has also been observed that performances of machine learning models measured in one domain have significant discrepancies to their performances in another distribution shifted domain (Hendrycks & Dietterich, 2019; Ovadia et al., 2019; Recht et al., 2019). Therefore, applying model selection techniques in the i.i.d. settings to the labeled source domain is suboptimal in the target domain.

In this work, we simultaneously address these critical aspects in UDA from a new perspective of predicting a group accuracy. Specifically, we partition predictions into a set of groups and then estimate the group accuracy—the average accuracy of predictions in a group. When the group accuracy estimate accurately represents the expected accuracy of a model for samples in the group (e.g., group 1 in Figure 1(a)), using the group accuracy estimate as prediction confidence induces a well-calibrated classifier. When the average of the group accuracy estimates matches the mean expected accuracy (e.g., two dotted lines in Figure 1(a) are close to each other), it becomes a good model selection criterion.



(a) (b)

Figure 1. Figure 1(a) illustrates ideal and failure cases of IW-GAE with nine data points (red diamonds) from three groups (gray boxes). Group 1 is desirable for model calibration where the group accuracy estimation (a blue rectangle) well represents the individual expected accuracies of samples in the group. Conversely, group accuracy estimation could inaccurately represent the individual accuracies in the group due to a high variance of accuracies within the group (group 2) and a high bias of the estimator (group 3). For model selection, we aim to match the mean group accuracy estimation (the blue dotted line as an average of blue rectangles) to the mean expected accuracy (the red dotted line as an average of red diamonds), which can be induced by accurate group accuracy estimations for each group. Figure 1(b) explains the idea of encouraging two estimators close to each other. The shaded area for the IW-based estimator is possible group accuracy estimations from different IWs. IW-GAE finds the IW minimizing the opt error for reducing the group accuracy estimation error.

To this end, we propose **importance weighted group accuracy estimation (IW-GAE)** that aims to find importance weights (IWs) that induce an accurate group accuracy estimator under the distribution shifts. Specifically, we define two estimators for the group accuracy in the source domain (MC-based and IW-based estimators in Figure 1(b)), where only one of them depends on the IW. Then, we formulate a novel optimization problem for finding the IW that makes the two estimators close to each other (reducing opt error in Figure 1(b)). Through theoretical analyses and several experiments, we show that the optimization process results in an accurate group accuracy estimator for the target domain (small quantity of interest in Figure 1(b)), improving model calibration and model selection performances.

Our contributions can be summarized as follows: 1) We show when and why considering group accuracy, instead of the accuracy for individual samples, is statistically favorable, which can simultaneously benefit model calibration and model selection with attractive properties; 2) We propose a novel optimization problem for IW estimation that directly reduces group accuracy estimation error in UDA with theoretical analyses; 3) On average, IW-GAE improves state-of-the-art performances by 22% in the model calibration task and 14% in the model selection task.

Notation and problem setup Let X $\mathbb{R}^r$ and $Y = [K] := f1; 2; \ldots; Kg$ be input and label spaces. Let $\hat{Y} : X \to [K]$ be the prediction function of a model and $Y(x)$ is a K-dimensional categorical random variable related to a label at $X = x$. When there is no ambiguity, we represent $Y(x)$ and $\hat{Y}(x)$ as Y and $\hat{Y}$ for brevity. We are given a labeled source dataset $D_S = f(x_i^{(S)}; y_i^{(S)})g_{i=1}^{N^{(S)}}$ sampled from $p_{S_{XY}}$ and an unlabeled target dataset $D_T = fx_i^{(T)}g_{i2[N^{(T)}]}$ sampled from p_{T_X} where $p_{S_{XY}}$ is a joint data generating distribution of the source domain and p_{T_X} is a marginal distribution of the target domain. We also denote $\mathbb{E}_p[\,]$ as the population expectation and $\hat{\mathbb{E}}_p[\,]$ as its empirical counterpart. For $p_{S_{XY}}$ and $p_{T_{XY}}$, we consider a covariate shift; i.e., $p_{S_X}(x) \neq p_{T_X}(x)$ but $p_{S_{YjX}}(yjx) = p_{T_{YjX}}(yjx)$ for $x \in X$. For the rest of the paper, we use the same notation p_S for marginal distribution p_{S_X} and joint distribution $p_{S_{XY}}$ when there is no ambiguity. However, we use the explicit notation for $p_{S_{YjX}}$ and $p_{T_{YjX}}$ to avoid confusion.

2. Group Accuracy Estimation for Model Calibration and Selection

We address model calibration and model selection tasks in UDA by estimating the group accuracy. Specifically, we construct M groups $fG_n g_{n2[M]}$ with some grouping function $l^{(g)} : X \to [M]$. Then, for each group G_n, we estimate the average accuracy of target domain samples in G_n defined as $_T(G_n) := \mathbb{E}_{p_T}[\mathbf{1}(Y(X) = \hat{Y}(X))jG_n]$. In the following, we first give a motivation for estimating the group accuracy, instead of an expected accuracy for each sample. Then, we explain how to construct groups and use the group accuracy estimates for simultaneously solving model calibration and selection tasks.

2.1. Motivation for Estimating the Group Accuracy

Suppose we are given samples $D := f(x_i; y_i) \in G_n g_{i2[N_n]}$ and a classifier f. Let $(x_i) := \mathbb{E}_{YjX=x_i}[\mathbf{1}(Y(x_i) = f(x_i))]$ be an expected accuracy of f at x_i, which is our goal to estimate. Then, due to realization of a single label at each point, the observed accuracy $\hat{}(x_i) := \mathbf{1}(y_i = f(x_i))$ is a random sample from the Bernoulli distribution with parameter (x_i) that has a variance of $^2_{x_i} = (x_i)(1 \quad (x_i))$. Note that this holds when $x_i \neq x_j$ for $i \neq j$, which is the case for most machine learning scenarios.

Under this setting, we show the sufficient condition that the maximum likelihood estimator (MLE) of the group accuracy

outperforms the MLE of the individual accuracy.

Proposition 2.1. *Let* $\hat{}^{(id)}$ *and* $\hat{}^{(gr)}$ *be MLEs of individual and group accuracies. Then,* $\hat{}^{(gr)}$ *has a lower expected mean-squared error than* $\hat{}^{(id)}$ *if*

$$\frac{1}{4}(\max_{x' \in G_n} \quad (x') \quad \min_{x' \in G_n} \quad (x'))^2 \quad \frac{N_n \quad 1}{N_n}^2 \qquad (1)$$

where $^2 = \frac{1}{N_n} \sum_{i \in [N_n]} \quad ^2_{x_i}$ *with* $^2_{x_i} = \quad (x_i)(1 \quad (x_i))$.

The proof is based on bias-variance decomposition and the Popoviciu's inequality (Popoviciu, 1965), which is given in Appendix A.1. In Proposition 2.1, (1) is the condition under which the group accuracy estimator achieves a lower mean-squared error than the individual accuracy estimator. Crucially, we can reduce $\max_{x \in G_n} \quad (x) \quad \min_{x \in G_n} \quad (x)$ through a careful group construction that we discuss in Section 2.2. Further, a sufficient condition for (1) tends to be loose (cf. Appendix A.2). Therefore, under the loose condition on the group construction, *the group accuracy estimator would be statistically more favorable*.

2.2. Group Assignment

As seen in Figure 1(a), it is important to construct groups such that the expected accuracy of samples in the same group has a low variance. Therefore, we group examples by the maximum value of the softmax output as in Guo et al. (2017) based on an observation that the maximum value of the softmax output is highly correlated with accuracy in UDA (Wang et al., 2020). In addition, we note the result that an overall scale of the maximum value of the softmax output significantly varies from one domain to another (Yu et al., 2022). Thus, we adjust the sharpness of the softmax output in the target domain by introducing a *learnable* temperature parameter $t \in T$ where T is a bounded interval in $\mathbb{R}_+$.

Equipped with these ideas, we first gather a set of prediction confidences under a temperature t, denoted as $C^{(t)} := \{m(x_S; 1) | (x_S, y_S) \in D_S\} \cup \{m(x_T; t) | x_T \in D_T\}$ where $m(x; t)$ is the maximum value of the softmax output under t ($t = 1$ recovers the standard softmax). Then, we construct the n-th confidence group under t for $n \in [M]$ by

$$G_n^{(t)} := \{x \in D_S | q(\tfrac{n-1}{M}; C^{(t)}) \quad m(x; 1) < q(\tfrac{n}{M}; C^{(t)})\} \cup \{x \in D_T | q(\tfrac{n-1}{M}; C^{(t)}) \quad m(x; t) < q(\tfrac{n}{M}; C^{(t)})\} \qquad (2)$$

where $q(\frac{n}{M}; C^{(t)})$ is the $\frac{n}{M}$-th quantile of $C^{(t)}$. For the rest of the paper, we let $I^{(g)}(x_S) = \min\{k \in [M] : m(x_S; 1) < q(\frac{k}{M}; C^{(t)})\}$ if $x_S \quad p_S$ and $I^{(g)}(x_T) = \min\{k \in [M] : m(x_T; t) < q(\frac{k}{M}; C^{(t)})\}$ if $x_T \quad p_T$.

2.3. Model Selection and Model Calibration

Before explaining how to obtain an accurate group accuracy estimator $\hat{}_T(G_n^{(t)})$ in Section 3, we first show how to use $\hat{}_T(G_n^{(t)})$ to simultaneously solve model calibration and model selection tasks with attractive properties.

Model calibration We use $\hat{}_T(G^{(t)}_{I^{(g)}(x)})$ as an estimate of prediction confidence on $x \quad p_T$. Then, we can address the model calibration task in UDA with a bounded calibration error. Specifically, an expected squared calibration error can be decomposed as the sum of the variance of the accuracies for samples in the same group and a squared group accuracy estimation error (cf. Proposition A.1); that is, $\mathbb{E}_{p_T}[(P(Y = \hat{Y}) \quad \hat{}_T(G^{(t)}_{I^{(g)}(X)}))^2] = \sum_{n \in [M]} \mathbb{E}_{p_T}[\mathbf{1}(X \in G_n^{(t)})] (Var(P(Y = \hat{Y}) | G_n^{(t)}) + (\quad _T(G_n^{(t)}) \quad \hat{}_T(G_n^{(t)}))^2)$. Thus, combined with the guarantees about the group accuracy estimation error (cf. Proposition 3.1 and (5)), our approach can enjoy the bounded calibration error unlike previous approaches using $m(x; t)$ as the prediction confidence estimate (Park et al., 2020; Wang et al., 2020).

Model selection We use the average group accuracy estimate $\mathbb{E}_{p_T}[\hat{}_T(G^{(t)}_{I^{(g)}(X)})]$ computed with a hold-out target domain dataset as the model selection criteria. Again, this criteria estimates the average accuracy of the model with a bounded error due to the Cauchy-Schwarz inequality; that is, $|\mathbb{E}_{p_T}[P(Y = \hat{Y})] \quad \mathbb{E}_{p_T}[\hat{}_T(G^{(t)}_{I^{(g)}(X)})]| \quad (\mathbb{E}_{p_T}[(P(Y = \hat{Y}) \quad \hat{}_T(G^{(t)}_{I^{(g)}(X)}))^2])^{1/2}$. In addition, compared to the approaches (Sugiyama et al., 2007; You et al., 2019) aiming to estimate only the mean accuracy in p_T, our approach will have an additional regularization effect from encouraging accurate group accuracy estimation for each group.

3. Accurate Group Accuracy Estimation via IW-GAE

In this section, we propose IW-GAE that aims to accurately estimate $\quad _T(G_n^{(t)})$ by using a novel idea tailored for UDA where $Y(x)$ is available for $x \quad p_S$ but not for $x \quad p_T$. A core idea behind IW-GAE is to use importance weighting, which is appealing due to its statistical exactness for dealing with two different probability distributions under the absolute continuity condition (Sugiyama et al., 2007). Specifically, we define the **target group accuracy** of a group $G_n^{(t)}$ with the true IW $w \quad (x) := \frac{p_T(x)}{p_S(x)}$ as

$$\quad _T(G_n^{(t)}; w \quad) = \mathbb{E}_{p_S}\left[w \quad (X)\mathbf{1}(Y = \hat{Y}) | G_n^{(t)} \right] \frac{P(X_S \in G_n^{(t)})}{P(X_T \in G_n^{(t)})} \qquad (3)$$

where X_S and X_T are random variables having densities p_S and p_T, respectively. We denote $\hat{}_T(G_n^{(t)}; w \quad)$ to be the expectation with respect to the empirical measure. We also define the **source group accuracy** of $G_n^{(t)}$ as

$$\quad _S(G_n^{(t)}; w \quad) = \mathbb{E}_{p_T}\left[\frac{\mathbf{1}(Y(X) = \hat{Y}(X))}{w \quad (X)} | G_n^{(t)} \right] \frac{P(X_T \in G_n^{(t)})}{P(X_S \in G_n^{(t)})}. \qquad (4)$$

(a) (b) (c)

Figure 2.Illustration of correlations between the optimization error and the source and target group accuracy estimation errors. Each point corresponds to a different IW estimator and the values are measured on the Of ceHome dataset (720 IW estimators in total). See Appendix G for more detailed discussions and analyses.

Given $_T(G_n^{(t)}; w)$, the group accuracy estimation problem can be reduced to the importance weight estimation problem; that is, nding w such that $_T(G_n^{(t)}; w) \approx _T(G_n^{(t)}; w)$. A typical approach for solving this estimation problem is to accurately approximate IW, i.e., $w \approx w$, which is challenging in high-dimensional spaces. In this work, we propose an optimization-based approach that minimizes the group accuracy estimation error during the IW estimation process, circumventing the dif culty of directly estimating w.

3.1. Motivation for an Optimization-Based Approach

Our idea for accurately estimating the "target" group accuracy with IW estimator w is to de ne two estimators for the "source" group accuracy de ned in (4), with one estimator dependent on w, and to encourage the two estimators to agree with each other. This approach can be validated because the target accuracy estimation error of w can be upper bounded by its source accuracy estimation error; that is,

$$|_T(G_n^{(t)}; w) - _T(G_n^{(t)}; \hat{w})| \le w_n^{(ub)} \, |_S(G_n^{(t)}; w) - _S(G_n^{(t)}; \hat{w})| \, \left(\frac{\hat{P}(X_T \in G_n^{(t)})}{\hat{P}(X_S \in G_n^{(t)})}\right)^2 \quad (5)$$

where $w_n^{(ub)} = \sup_{x \in \mathrm{Supp}(p_T(\cdot|G_n^{(t)}))} w(x)$ and the bound is tight when $\hat{w}(x) = w_n^{(ub)}$ for all $x \in \mathrm{Supp}(p_T(\cdot|G_n^{(t)}))$.

Based on the fact that labeled samples are available in the source domain, we de ne a rst estimator of $_S(G_n^{(t)}; w)$ with the simple Monte-Carlo estimation by

$$\hat{}_S^{(MC)}(G_n^{(t)}) = \hat{E}_{p_S}[1(Y = \hat{Y}) | G_n^{(t)}]. \quad (6)$$

We note that $\hat{}_S^{(MC)}(G_n^{(t)})$ serves as a guide for the agreement between two estimators because it accurately estimates $_S(G_n^{(t)}; w)$ with a small error of $O(1/\sqrt{|G_n^{(t)}(D_S)|})$ where $G_n^{(t)}(D_S) := \{(x_k, y_k) \in D_S : x_k \in G_n^{(t)}\}$.

Based on the fact that input features are available both in source and target domains, we de ne a second estimator of $_S(G_n^{(t)}; w)$ with importance weighting. Speci cally, by assuming $E_{p_{T_{Y|x}}}[1(Y(x) = \hat{Y}(x))] = _T(G_n^{(t)}; w)$ for all $x \in G_n^{(t)}$ and replacing w by $\hat{w}$ in (4), we obtain the second estimator as a function of $\hat{w}$, which is given by

$$\hat{}_S^{(IW)}(G_n^{(t)}; \hat{w}) := \frac{\hat{P}(X_T \in G_n^{(t)})}{\hat{P}(X_S \in G_n^{(t)})} \hat{E}_{p_T}\left[\frac{\hat{}_T(G_n^{(t)}; \hat{w})}{\hat{w}(X)} | G_n^{(t)}\right]$$
$$= \hat{E}_{p_T}\left[\frac{1}{\hat{w}(X)} | G_n^{(t)}\right] \hat{E}_{p_S}[1(Y = \hat{Y}) \hat{w}(X) | G_n^{(t)}] \quad (7)$$

where $\hat{}_T(G_n^{(t)}; \hat{w})$ is an empirical estimate of the target accuracy de ned in (3), $\hat{P}(X_T \in G_n^{(t)}) := \hat{E}_{p_T}[1(X \in G_n^{(t)})]$, and $\hat{P}(X_S \in G_n^{(t)}) := \hat{E}_{p_S}[1(X \in G_n^{(t)})]$.

Crucially, two estimators, $\hat{}_S^{(MC)}(G_n^{(t)})$ and $\hat{}_S^{(IW)}(G_n^{(t)}; \hat{w})$, can be similar to each other if the target group accuracy estimation with $\hat{w}$ is accurate (cf. (7)); that is, $\hat{}_T(G_n^{(t)}; \hat{w}) \approx _T(G_n^{(t)}; w)$. Therefore, by encouraging consistency between two estimators via solving the optimization problem developed in Section 3.2, IW-GAE can accurately estimate the group accuracy not only in the source domain but also in the target domain. This conceptual attractiveness is empirically veri ed in Figure 2(a), which compares group accuracy estimation errors in the source and target domains from 720 IWs found by IW-GAE. Speci cally, we found that under the optimal IW found by IW-GAE, group accuracy estimation errors in the source and target domains, $|_S(G_n^{(t)}; w) - _S(G_n^{(t)}; \hat{w})|$ and $|_T(G_n^{(t)}; w) - _T(G_n^{(t)}; \hat{w})|$ in (5), are strongly correlated with a high Pearson correlation coef cient of 0.8.

3.2. Formulating an Optimization Problem

In this section, we aim to solve an optimization problem such that $\min_{t,\hat{w}} (\hat{}_S^{(IW)}(G_n^{(t)}; \hat{w}) - \hat{}_S^{(MC)}(G_n^{(t)}))^2$. Unfortunately, $\hat{}_S^{(IW)}(G_n^{(t)}; \hat{w})$ in (7) is non-convex with respect to $\hat{w}$ and non-smooth with respect to t, which is in general not effectively solvable with optimization methods (Jain et al., 2017). Further, directly solving an optimization problem in the function space of $\hat{w}$ or optimizing over IW values for each $x \in X$ would be computationally demanding. Therefore, we introduce the following techniques to formulate the optimization problem in a tractable way.

Relaxed reformulation We separately estimate IWs for the source and target domains, denoted as $\hat{w}^{(S)}$ and $\hat{w}^{(T)}$, and then encourage their agreement through constraints. As a result, the estimator in (7) becomes $\hat{\alpha}_S^{(IW)}(G_n^{(t)}; \hat{w}^{(S)}, \hat{w}^{(T)}) = \hat{E}_{p_T}[\frac{1}{\hat{w}^{(T)}(X)} | G_n^{(t)}] \hat{E}_{p_S}[\mathbb{1}(Y = \hat{Y}) \hat{w}^{(S)}(X) | G_n^{(t)}]$, which is coordinate-wise convex.

Discretize $\mathcal{T}$ We use a discrete set $\mathcal{T} := \{t_1, t_2, \ldots, t_n\}$ based on the facts that a group separation is not sensitive to small changes in t and the inner optimization is not smooth with respect to t. We remark that the inner optimization problem with respect to $w^{(S)}$ and $w^{(T)}$ is readily solvable, so the discrete optimization over $\mathcal{T}$ can be performed without much computational overhead.

Binned IWs We approximate $w^{(S)}$ and $w^{(T)}$ by the binned IWs. Specifically, $\mathcal{X}$ is partitioned into B number of bins: $\mathcal{X} = [_{i=1}^{B} \mathcal{B}_i$ where $\mathcal{B}_i = \{x \in \mathcal{X} \mid I^{(B)}(x) = i\}$ and $I^{(B)} : \mathcal{X} \to [B]$. Then, we assign the same IW value to all samples in the same bin; that is, $w^{(S)}(x_S) = w_j^{(S)}$ for $x_S \in \mathcal{B}_j \cap \mathcal{D}_S$ and $w^{(T)}(x_T) = w_j^{(T)}$ for $x_T \in \mathcal{B}_j \cap \mathcal{D}_T$. In this way, the number of decision variables in the inner optimization problem is reduced to $2B$. Further, we incorporate the recently proposed confidence interval (CI) estimation method for the binned IWs (Park et al., 2022) into the constraints. We denote $\mathcal{I}_j$ be the CI of the true binned IW of $\mathcal{B}_j$, which is obtained by applying the Clopper-Pearson CI (Clopper & Pearson, 1934) (cf. Appendix C.1). We let $w^{(S)} = (w_1^{(S)}, \ldots, w_B^{(S)})$ and $w^{(S)}(x) = w_{I^{(B)}(x)}^{(S)}$. We also define $w^{(T)}$ and $w^{(T)}(x)$ in the same way.

Assembling the three techniques, we can effectively solve the group accuracy estimation problem by finding binned IWs $w^y(n; t^y) \in \mathbb{R}_+^{2B}$ for $n \in [M]$ by solving the following nested optimization (see Algorithm 1 for pseudocode) $t^y \in \arg\min_{t \in \mathcal{T}} \sum_{n \in [M]} (\hat{\alpha}_S^{(MC)}(G_n^{(t)}) - \hat{\alpha}_S^{(IW)}(G_n^{(t)}; w^y(n; t)))^2$ where $w^y(n; t)$ is a solution of

$$\min_{w^{(S)}, w^{(T)}} (\hat{\alpha}_S^{(MC)}(G_n^{(t)}) - \hat{\alpha}_S^{(IW)}(G_n^{(t)}; w^{(S)}, w^{(T)}))^2 \quad (8)$$

$$\text{s.t. } w_i^{(S)} \in \mathcal{I}_i, \quad \text{for } i \in [B] \quad (9)$$

$$w_i^{(T)} \in \mathcal{I}_i, \quad \text{for } i \in [B] \quad (10)$$

$$\| w_i^{(T)} \mathcal{D}_T - w_i^{(S)} \mathcal{D}_S \|_2^2 \le \epsilon_{tol}, \quad \text{for } i \in [B] \quad (11)$$

$$\hat{E}_{p_S}[w^{(S)}(X) | G_n^{(t)}] \frac{\hat{P}(X_T \in G_n^{(t)})}{\hat{P}(X_S \in G_n^{(t)})} \ge \epsilon_{pr} \quad (12)$$

$$\hat{E}_{p_T}[\frac{1}{\hat{w}^{(T)}(X)} | G_n^{(t)}] \frac{\hat{P}(X_S \in G_n^{(t)})}{\hat{P}(X_T \in G_n^{(t)})} \ge \epsilon_{pr} \quad (13)$$

where ϵ_{tol} and ϵ_{pr} are small constants. Box constraints (9) and (10) ensure that the obtained solution is in the CI which bounds the estimation error of $w_i^{(S)}$ and $w_i^{(T)}$ by $|\mathcal{I}_i|$ and guarantees their asymptotic convergences to the true binned IW (Thulin, 2014). This can also bound the target group accuracy estimation error $|\epsilon_S(G_n^{(t)}; w^{(S)}) - \epsilon_T(G_n^{(t)}; w^{(S)})| \le \max_{b \in [B]} |\mathcal{I}_b| P(X_S \in G_n^{(t)}) / P(X_T \in G_n^{(t)})$. Constraint (11) corresponds to the relaxation for removing non-convexity of the original objective, and $\epsilon_{tol} = 0$ recovers the original objective. Constraints (12) and (13) are based on the equalities that the true IW $w^*(\cdot)$ satisfies: $E_{p_S}[w^*(X) | X \in G_n^{(t)}] = \frac{P(X_T \in G_n^{(t)})}{P(X_S \in G_n^{(t)})}$ and $E_{p_T}[1/w^*(X) | X \in G_n^{(t)}] = \frac{P(X_S \in G_n^{(t)})}{P(X_T \in G_n^{(t)})}$. After solving the optimization, we use the first B elements of $w^y(n; t^y)$ that correspond to the optimal $w^{(S)}$ for estimating group accuracy of G_n, which is denoted by $w^y(n)$.

3.3. Analyzing the Optimization Problem

The optimization problem in (8)-(13) aims to estimate the truncated IW $w^*(x | G_n^{(t)}) := \frac{p_T(x | G_n^{(t)})}{p_S(x | G_n^{(t)})}$ for each $G_n^{(t)}$ that can induce an accurate source group accuracy estimator. However, the astute reader might notice that the objective in (8) does not measure the source group accuracy estimation error. In the following proposition, we show that solving the optimization problem minimizes the upper bound of the source group accuracy estimation error, thereby the target group accuracy estimation error due to (5).

Proposition 3.1. Let $w^y(n)$ be a solution to the nested optimization problem with $\epsilon_{tol} = 0$ and $\epsilon_{pr} = 0$. Let $\epsilon_{opt}(w^y(n)) := (\hat{\alpha}_S^{(MC)}(G_n^{(t)}) - \hat{\alpha}_S^{(IW)}(G_n^{(t)}; w^y(n)))^2$ be the objective value. For $\tilde{\epsilon} > 0$, the following inequality holds with probability at least $1 - \tilde{\epsilon}$:

$$|\epsilon_S(G_n^{(t)}; w^*) - \epsilon_S(G_n^{(t)}; w^y(n))| \quad (14)$$

$$\le \epsilon_{opt}(w^y(n)) + \epsilon_{stat} + \mathrm{IdentBias}(w^y(n); G_n^{(t)}) \quad (15)$$

where $\epsilon_{stat} \in O(\sqrt{\log(1/\tilde{\epsilon}) / |G_n^{(t)}(D_S)|})$ for $w^y(n) := \min_{i \in [B]} \{w_i^y(n)\}$ and $\mathrm{IdentBias}(w^y(n); G_n^{(t)}) = \frac{P(X_T \in G_n^{(t)})}{2P(X_S \in G_n^{(t)})} (E_{p_T}[(\mathbb{1}(Y = \hat{Y}) - \hat{\alpha}_T(G_n^{(t)}; w^y(n)))^2 | G_n^{(t)}] + \frac{1}{w^y(n)^2})$.

The proof is based on the Cauchy-Schwarz inequality, which is provided in Appendix A.4. Proposition 3.1 shows that we can reduce the source accuracy estimation error by reducing $\epsilon_{opt}(w^y(n))$ by solving the optimization problem. Here, we note that a large value of $\epsilon_{stat} + \mathrm{IdentBias}(w^y(n); G_n^{(t)})$ or a looseness of (15) could significantly decrease the effectiveness of IW-GAE. However, in the empirical analyses presented in Figures 2(b) and 2(c), it turns out that reducing $\epsilon_{opt}(w^y(n))$ can effectively reduce the group accuracy estimation in both source and target domains. Finally, we note that $\mathrm{IdentBias}(w^y(n); G_n)$ can be reduced by decreasing the variance of the correctness within the group, which advocates our group construction with the maximum value of the softmax output (cf. Proposition A.2).

Table 1. Model calibration benchmark results of MDD (Of ceHome) and CDAN (DomainNet and VisDa-2017). The numbers indicate the mean ECE across ten repetitions with boldface for the minimum mean ECE. Due to space limitations, we present the first six domain pairs of Of ceHome and DomainNet in the main body and the rest of them in Tables A1 and A2, respectively. However, we report average performance among all pairs in Avg*. Oracle is obtained by applying TS with labeled test samples in the target domain.

	Of ceHome							DomainNet							VisDa-2017
Method	Ar-Cl	Ar-Pr	Ar-Rw	Cl-Ar	Cl-Pr	Cl-Rw	Avg*	Cl-Pt	Cl-Rw	Cl-Sk	Pt-Cl	Pt-Rw	Pt-Sk	Avg*	Sim-Rw
Vanilla	40.61	25.62	15.56	33.83	25.34	24.75	24.37	13.23	6.36	12.92	9.75	6.35	15.56	10.06	21.63
TS	35.86	22.84	10.60	28.24	20.74	20.06	24.01	12.95	5.95	13.32	6.40	3.90	11.07	9.22	22.42
CPCS	22.93	22.07	10.19	26.88	18.36	14.05	19.79	5.64	21.90	7.70	5.14	7.72	7.90	9.60	22.42
IW-TS	32.63	22.90	11.27	28.05	19.65	18.67	23.26	16.76	16.70	12.53	5.29	7.84	4.34	10.49	22.19
TransCal	33.57	20.27	8.88	26.36	18.81	18.42	20.84	18.51	29.63	20.92	23.02	31.83	17.58	25.39	18.79
IW-GAE	12.78	4.70	12.93	7.52	4.42	4.11	8.93	6.06	8.15	5.38	7.45	3.89	3.94	6.32	14.70
Oracle	10.45	10.72	6.47	8.10	7.62	6.55	8.42	4.55	2.78	4.01	3.10	3.72	2.72	3.02	5.48

4. Experiments

In this section, we extensively evaluate IW-GAE on model calibration and selection tasks. Since both tasks are based on UDA classification tasks, we first provide the common setup and task-specific setup such as the baselines and evaluation metrics in the corresponding sections.

Datasets We use Of ceHome (Venkateswara et al., 2017) containing around 15,000 images of 65 categories from four domains (art, clipart, product, real-world), VisDA-2017 (Peng et al., 2017) containing around 280,000 images of 12 categories from two domains (real and synthetic images), and DomainNet (Peng et al., 2019) containing around 570,000 images of 345 categories from six domain pairs (clipart, real, sketch, infograph, painting, quickdraw).

Base models We consider maximum mean discrepancy (MDD; (Zhang et al., 2019)), conditional domain adversarial network (CDAN; (Long et al., 2018)), and maximum classifier discrepancy (MCD; (Saito et al., 2018)) with ResNet-50 (He et al., 2016) as the backbone neural network, which are the most popular high-performing UDA methods. Details about the training configurations are given in Appendix D.

IW-GAE implementation details We solve the optimization problem in (8)-(13) by sequential least square programming (Kraft, 1988) because it is a constrained nonlinear optimization problem with box constraints. Also, we set the number of groups $M = 10$ and the number of bins $B = 10$ for all experiments, which are taken from the standard range $[10, 20]$ used for binning samples based on summary statistics (Guo et al., 2017; Park et al., 2022). Finally, we set $T = \{0.85, 0.90, 0.95, 1.00, 1.05, 1.10\}$. We provide further details in Appendix D.

4.1. Model Calibration Performance

Setup & Metric In this experiment, our goal is to match the confidence of a prediction to its expected accuracy in the target domain. Following the standard (Guo et al., 2017; Park et al., 2020; Wang et al., 2020), we use expected calibration error (ECE) on the test dataset as a measure of calibration performance. The ECE measures the average absolute difference between the confidence and accuracy of binned groups, which is defined as

$$\mathrm{ECE}(D_T) = \sum_{n \in [E]} \frac{|M_n|}{|D_T|} |\widehat{\mathrm{Acc}}(M_n) - \widehat{\mathrm{Conf}}(M_n)| \quad (16)$$

where $M_n := \{x_i \in D_T \mid \frac{n-1}{E} < m(x_i; 1) \leq \frac{n}{E}\}$, $\widehat{\mathrm{Acc}}(M_n)$ is the average accuracy in M_n, and $\widehat{\mathrm{Conf}}(M_n)$ is the average confidence in M_n. For IW-GAE, $\widehat{\mathrm{Conf}}(M_n) = \frac{1}{|M_n|} \sum_{x \in M_n} \hat{\alpha}_T(G^{(t^y)}_{l^{(g)}(x)}; w^y(l^{(g)}(x)))$, which is the average of group accuracy estimations that each $x \in M_n$ belongs to. We use $E = 15$ following the standard value (Guo et al., 2017; Wang et al., 2020).

Baselines We consider the following five different baselines: The *vanilla* method uses a maximum value of the softmax output as the confidence of the prediction. We also consider temperature scaling-based methods that adjust the temperature parameter by maximizing the following calibration measures: Temperature scaling (TS; (Guo et al., 2017)): the log-likelihood on the source validation dataset; IW temperature scaling (IW-TS): the log-likelihood on the importance weighted source validation dataset; Calibrated prediction with covariate shift (CPCS): the Brier score (Brier, 1950) on the importance weighted source validation dataset; TransCal (Wang et al., 2020): the ECE on the importance weighted source validation dataset with a bias and variance reduction technique. These methods also use a maximum value of the (temperature-scaled) softmax output as the confidence. For methods with the IW, we use a logistic regression-based IW estimator as in Wang et al. (2020) (cf. Appendix C.2).

Results As shown in Table 1, IW-GAE achieves the best average ECEs across different base models and datasets. Specifically, IW-GAE outperforms state-of-the-art performances by 53% on Of ceHome, 31% on DomainNet, and 21% on VisDa-2017. Further, in additional experiments with different base methods (CDAN and MCD) on Of ceHome, IW-GAE consistently outperforms state-of-the-art

Table 2.Checkpoint selection benchmark results of MDD with ResNet-50 on Of ceHome. The numbers indicate the mean test accuracy of selected model across ten repetitions with boldface for the maximum mean test accuracy. Due to space limitations, we present the rst six domain pairs in the main body and the rest of them in Tables A3 and A4. However, we report average performance among all pairs in Avg*. Here, we also present two best methods among the target-only validation methods and the rest of them in Tables A3 and A4. Lower bound and Oracle indicate the accuracy of the models with the worst and best test accuracy, respectively.

	Hyperparameter Selection							Checkpoint Selection						
Method	Ar-Cl	Ar-Pr	Ar-Rw	Cl-Ar	Cl-Pr	Cl-Rw	Avg*	Ar-Cl	Ar-Pr	Ar-Rw	Cl-Ar	Cl-Pr	Cl-Rw	Avg*
Vanilla	53.31	70.96	77.44	59.70	65.17	69.96	65.45	47.22	74.14	77.76	61.85	70.96	71.59	67.47
IWCV	53.24	69.61	72.50	59.70	65.17	67.50	65.18	54.46	74.22	72.27	61.48	70.49	70.62	67.48
DEV	53.31	70.72	77.44	59.79	67.99	69.96	66.00	54.04	73.94	78.16	61.52	63.19	70.70	67.39
InfoMax	54.34	70.96	77.53	61.48	69.93	71.06	67.79	54.32	74.72	77.90	62.79	71.03	71.47	68.38
TransScore	54.34	70.96	77.53	61.48	69.93	71.06	67.87	54.79	74.14	77.77	61.76	70.97	71.48	68.38
IW-GAE	54.34	70.96	78.47	61.48	69.93	71.06	67.95	54.32	73.98	78.51	61.96	71.25	71.70	68.48
Lower bound	52.51	69.27	72.50	59.70	65.17	67.50	64.10	41.90	64.88	72.27	52.00	58.48	62.13	58.21
Oracle	54.34	70.96	78.47	61.48	69.93	71.06	68.01	54.80	74.79	78.61	62.79	71.59	72.18	68.95

Table 3.Additional model calibration benchmark with post-hoc calibration methods under unknown distribution shifts. Due to space limitations, we present the rst six domain pairs in the main body and the rest of them in Table A1.

Method	Ar-Cl	Ar-Pr	Ar-Rw	Cl-Ar	Cl-Pr	Cl-Rw	Avg*
Vanilla	40.61	25.62	15.56	33.83	25.34	24.75	27.37
PTS	31.91	24.36	10.65	22.81	20.42	15.92	21.91
AvUTS	29.59	25.55	10.40	31.81	26.06	26.15	28.17
TransCal	33.57	20.27	8.88	26.36	18.81	18.42	20.84
IW-GAE	12.78	4.70	12.93	7.52	4.42	4.11	8.93

performances by 2% and 5%, respectively (cf. Tables A5 and A6). Given that the second best model varies for a different dataset and a different base model, we believe that the consistent improvements by IW-GAE indicate its signi cant robustness compared to the baselines.

In Table 3, we also compare IW-GAE with two recent post-hoc calibration methods, called PTS (Tomani et al., 2022) and AvUTS (Krishnan & Tickoo, 2020), which are designed to perform model calibration without using target domain samples. As post-hoc calibration methods under general distribution shifts, PTS and AvUTS show compatible results with IW-GAE and outperform some baselines in some domain pairs, i.e., certain types of distribution shifts. However, they cannot achieve better average performances than baselines or IW-GAE that explicitly consider the target distribution shift through importance weighting. The results show the advantage of explicitly using the information about the distribution shifts for the model calibration task in UDA.

4.2. Model Selection Performance

Setup & Metric In this experiment, we perform two important model selection tasks of choosing the best checkpoint and the best hyperparameter. Speci cally, for the checkpoint selection, we train MDD on the Of ceHome dataset for 30 epochs and save the checkpoint at the end of each epoch. For the hyperparameter selection, we repeat training the MDD method by changing its key hyperparameter of margin co ef cient from 1 to 8 (the default value is 4). Given a set of models from different checkpoints or different hyperparameters, we choose the best model based on a model selection criterion (such as IW-GAE or other baselines). Speci cally, for IW-GAE, we choose the model with the maximum value of the mean group accuracy estimations, which is computed by $\frac{1}{|D_T|}\sum_{x \in D_T} \hat{}_T(G^{(\hat{y})}_{I^{(g)}(x)}; w^{\hat{y}}(I^{(g)}(x)))$. Then, we compare the test target accuracy of the chosen models under different model selection methods.

Baselines We consider the following baselines that evaluate the model's performance in terms of the following criterion: Vanilla: the minimum classi cation error on the source validation dataset; importance weighted cross validation (IWCV) (Sugiyama et al., 2007): the minimum importance-weighted classi cation error on the source validation dataset; Deep embedded validation (DEV) (You et al., 2019): the minimum deep embedded validation risk on the source validation dataset; Target-only validation methods: the maximum value of some pre-de ned measures, e.g., the average negative entropy of predictions, on the unlabeled target validation dataset. Again, we use a logistic regression-based IW estimator for methods with the IW (IWCV and DEV). For the target-only validation methods, we consider InfoMax (Shi & Sha, 2012), Corr-C (Tu et al., 2023), SND (Saito et al., 2021), MixVal (Hu et al., 2024), and TransScore (Yang et al., 2024).

Results Table 2 shows that model selection with IW-GAE achieves the best average accuracy among IW-based model selection methods, improving state-of-the-art by 9% in the checkpoint selection 18% in the hyperparameter selection in terms of the relative scale of lower and upper bounds of accuracy. Note that IWCV does not improve the vanilla method on average, which could be due to the inaccurate IW estimation by the logistic regression-based method. In this sense, IW-GAE has the advantage of depending less on the performance of the IW estimator since the estimated value is used to construct bins for the CI, and then the exact value is found by solving the separate optimization problem.

Table 4.Results of an ablation study with four randomly selected domain pairs in Of ceHome. The numbers indicate the mean ECE of MDD with ResNet-50.

Method	Ar-Pr	Pr-Cl	Rw-Cl	Rw-Pr	Avg
Vanilla	40.61	38.62	36.51	14.01	32.44
CPCS	22.07	29.20	26.54	11.14	22.24
TransCal	20.27	29.86	29.90	10.00	22.51
Grouping by IW	14.18	34.67	35.08	5.30	22.31
W/O CI	11.00	29.73	24.44	2.09	16.82
IW-Mid	31.62	30.35	26.32	10.60	24.70
IW-GAE	4.70	17.49	9.52	8.14	9.97

In addition, we observe that all target-only validation methods except SND outperform the IW-based baselines (cf. Table 2). The results are consistent with the recent empirical observations that target-only validation methods are more favorable than IW-based methods for the model selection task in UDA (Saito et al., 2021; Hu et al., 2024). However, notably, they underperform IW-GAE in both checkpoint and hyperparameter selection tasks, which strongly supports the practical advantage of IW-GAE as a model selection method. We believe that the impressive empirical performance of IW-GAE could bring more attention to the IW-based model selection techniques in the machine learning community, which is a principled statistical method for the model selection task under distribution shifts but has been considered impractical and less effective.

4.3. Ablation Study

We conduct an ablation study by performing the model calibration task without key components of IW-GAE.

Group construction with the softmax output We first examine the effectiveness of group construction based on the maximum value of the softmax output by examining a group construction function based on IW. In Table 4, we can see that grouping by the IW significantly reduces the performance of IW-GAE due to a large variance of prediction accuracy within a group (cf. the case of group 2 in Figure 1(a)). Specifically, the large value of $\mathrm{Var}(\mathbb{1}(Y = \hat{Y}) \mid G_n^{(t)})$ increases the $\mathrm{IdentBias}(w^y(n); G_n^{(t)})$, which can loosen the upper bound of the source group accuracy estimation error in (15). Therefore, it is important to construct groups so that each group has a low variance of the correctness of predictions, as our design developed in Section 2.2.

Constraints from the CI estimation method Next, we examine the dependency of the effectiveness of IW-GAE on the CI estimation method (Park et al., 2022) by setting only minimum and maximum values of IWs (W/O CI); that is, $_i = [1/6; 6.0]$ for $i \in [M]$. In Table 4, we can see that IW-GAE outperforms strong baseline methods (CPCS and TransCal) even under this naive interval of IWs. However, the performance is reduced compared to the setting with the sophisticated CI estimator. In this regard, developing an optimization-based IW estimation method that works effectively without the CI estimator in the constraints could be an interesting future direction.

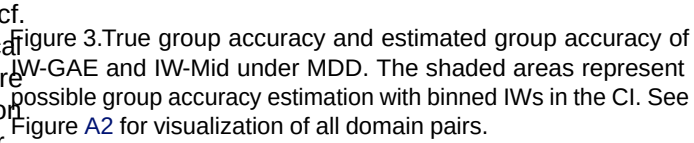


Figure 3.True group accuracy and estimated group accuracy of IW-GAE and IW-Mid under MDD. The shaded areas represent possible group accuracy estimation with binned IWs in the CI. See Figure A2 for visualization of all domain pairs.

IW optimization To further show the effectiveness of the optimization in IW-GAE, we also test the method of selecting the middle point in the CI proposed in Park et al. (2022) as an IW estimator (IW-Mid), which originates herein. For IW-Mid, we perform both model calibration and selection tasks across different base models and datasets. Surprisingly, IW-Mid achieves the better average performances than other IW-based baselines in some benchmarks (cf. Tables A1-A4). However, its performance is worse and significantly unstable compared to IW-GAE, such as achieving the worse performance even than the vanilla method in experiments with CDAN (Table A5) and MCD (Table A6). This means that the CI estimation method does not effectively estimate the group accuracy without properly selecting the exact IW through our optimization method, which is consistent with the theoretical result in Proposition 3.1. The instability and inaccuracy of IW-Mid compared to IW-GAE also can be identified in the qualitative evaluation of their group accuracy estimations (cf. Figure 3), which shows that the true accuracy is close to IW-Mid only in some cases.

4.4. Sensitivity Analysis

We perform a sensitivity analysis with respect to key hyperparameters of the number of accuracy groups $M \in [2; 27]$ and the number of bins $B \in [4; 27]$ (the default value for both M and B is 10). In Figure 4, note that the average performance changes under different hyperparameter values are somewhat stable; the average changes are within the range of 10% for most cases, even though a large variance appears for extreme values such as $M = 4$ and $B = 27$. Therefore, the results show that IW-GAE would consistently outperform state-of-the-art methods under changes in

(a) (b)

Figure 4. Sensitivity analysis with respect to M (a) and B (b) on four domain pairs (Ar-Pr, Pr-Cl, Rw-Cl, Rw-Pr) in Office-Home. The shaded areas represent areas between the minimum and the maximum changes in ECE (lower is better).

and B within their standard range $[10, 20]$ since the best baseline (TransCal) achieves the mean ECE score about 40% higher for the selected domain pairs (cf. Table 1).

5. Related Work

Model calibration in UDA Although post-hoc calibration methods (Guo et al., 2017) and Bayesian methods (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Snoek et al., 2018) have been achieving impressive calibration performances in the i.i.d. setting, it has been shown that most of the calibration improvement methods fall short under distribution shifts (Ovadia et al., 2019) (see Appendix B.1 for more discussion). While there have been attempts to perform model calibration under unknown distribution shifts by simulating the distribution shifts (Salvador et al., 2021), designing a robust loss function that prevents extrapolations with high confidences (Krishnan & Tickoo, 2020; Hebbalaguppe et al., 2022; Liu et al., 2022), and learning an instance-wise temperature parameter (Tomani et al., 2022), handling model calibration problems under general distribution shifts is challenging. However, the availability of unlabeled samples in the distribution shifted target domain relaxes the difficulty of model calibration in UDA. In particular, unlabeled samples in the target domain enable an IW formulation for the quantity of interests in the shifted domain. Therefore, the post-hoc calibration methods (e.g. (Guo et al., 2017)) can be applied by reweighting calibration measures such as the expected calibration error (Wang et al., 2020) and the Brier score (Park et al., 2020) in the source dataset with an IW. However, it remains unclear how the IW estimation error impacts the calibration error. Our approach, by contrast, can directly minimize the calibration error during the IW estimation process.

Model selection in UDA A standard procedure for model selection in the i.i.d. settings is the cross-validation, which enjoys statistical guarantees about bias and variance of model performance (Stone, 1977; Kohavi et al., 1995; Efron & Tibshirani, 1997). However, in UDA, the distribution shifts violate assumptions for the statistical guarantees. Furthermore, in practice, the accuracy measured in one domain is significantly changed in the face of natural/adversarial distribution shifts (Goodfellow et al., 2015; Hendrycks & Dietterich, 2019; Ovadia et al., 2019). To tackle the distribution shift problem, importance weighted cross validation (Sugiyama et al., 2007) applies importance sampling for obtaining an unbiased estimate of model performance in the distribution shifted target domain. Further, recent work in UDA controls variance of the importance-weighted cross validation with a control variate (You et al., 2019). These methods aim to accurately estimate the IW and then use an IW formula for the expected accuracy estimation. In this work, our method concerns the accuracy estimation error in the target domain during the process of IW estimation, which can potentially induce an IW estimation error but result in an accurate accuracy estimator. Finally, we remark a direction that aims to evaluate the model performance without source domain data based on neighborhood structure (Saito et al., 2021; Hu et al., 2024), prediction uncertainty in the target domain (Musgrave et al., 2022; Tu et al., 2023), and newly developed metrics (Yang et al., 2024).

6. Conclusion

In this work, we address the model calibration and selection tasks in UDA by estimating group accuracy, which is accurately estimated by solving a novel optimization problem. Specifically, we define a Monte-Carlo estimator and an IW-based estimator of group accuracy in the source domain. Then, we formulate an optimization problem that aims to find the IW making the two estimators close to each other. Crucially, the optimal IW provably leads to an accurate group accuracy estimator in the target domain. Our method achieves the best performances in both model calibration and selection tasks in UDA across a wide range of benchmark problems. We believe that the impressive performance gains by our method show a promising future direction of the (importance-weighted) group accuracy estimation for addressing critical challenges in UDA.

Limitations and future directions We note that all IW-based methods (CPCS, IW-TS, TransCal, IW-GAE) fail to improve the standard method in the i.i.d. scenario in our experiments with pre-trained large-language models (XLM-R (Conneau et al., 2019) and GPT-2 (Solaiman et al., 2019)). We conjecture that these models are less subject to the distribution shifts due to massive amounts of training data that may include the target domain datasets, so applying the methods in the i.i.d. setting can work effectively. In this regard, we leave the following important research questions: "Are IW-based methods less effective under mild distribution shifts?" and "Can we develop methods effective for all levels of distribution shifts?"

Impact Statement

In this work, we consider model calibration and model selection problems under distribution shifts, which hold great signi cance in practice, especially in safety-critical domains such as medical diagnosis and autonomous driving. Specifically, the well-calibrated models can properly require human intervention for uncertain predictions, preventing any catastrophic consequences from overcon dent predictions in automated systems. In addition, an accurate model selection allows the deployment of high-performing models in the distribution-shifted environment. Further, precisely evaluating model performance enables practitioners to reject deploying the model if its estimated performance is below a certain acceptable level. As evidenced in our extensive evaluations, IW-GAE helps to obtain robust and trustworthy models equipped with these ideal properties in unsupervised domain adaptation settings. Also, the core idea behind IW-GAE is to introduce a notion of group accuracy and then estimate it by optimizing the importance weight, which does not depend on particular characteristics or attributes of datasets. Therefore, IW-GAE would not leverage any biases inherent in the datasets, which prevents our work from having potential negative societal consequences. To sum up, we believe in a positive broader impact of IW-GAE.

Acknowledgement

We would like to thank Jihyeon Hyeong, Yuchen Lou, Jiezhong Wu, and anonymous reviewers for insightful discussions and helpful suggestions in writing the manuscript.

References

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning for differing training and test distributions. In *International Conference on Machine Learning*, 2007.

Brier, G. W. Veri cation of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Cai, T., Gao, R., Lee, J., and Lei, Q. A theory of label propagation for subpopulation shift. In *International Conference on Machine Learning*, 2021.

Chen, Y., Wei, C., Kumar, A., and Ma, T. Self-training avoids using spurious features under domain shift. In *Advances in Neural Information Processing Systems*, 2020.

Clopper, C. J. and Pearson, E. S. The use of con dence or ducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Dawid, A. P. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Ebrahimi, S., Elhoseiny, M., Darrell, T., and Rohrbach, M. Uncertainty-guided continual learning with Bayesian neural networks. In *International Conference on Learning Representations*, 2020.

Efron, B. and Tibshirani, R. Improvements on cross-validation: The 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.

Gal, Y., Hron, J., and Kendall, A. Concrete dropout. *Advances in Neural Information Processing Systems*, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Hebbalaguppe, R., Prakash, J., Madan, N., and Arora, C. A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Hu, D., Liang, J., Liew, J. H., Xue, C., Bai, S., and Wang, X. Mixed samples as probes for unsupervised model selection in domain adaptation. *Advances in Neural Information Processing Systems*, 2024.

Jain, P., Kar, P., et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–363, 2017.

Jiang, J., Chen, B., Fu, B., and Long, M. Transfer-learning-library. https://github.com/thuml/Transfer-Learning-Library, 2020.

Joo, T., Chung, U., and Seo, M.-G. Being Bayesian about categorical probability. In International Conference on Machine Learning, 2020.

Kohavi, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In International Joint Conferences on Artificial Intelligence, 1995.

Kraft, D. A software package for sequential quadratic programming. Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt, 1988.

Krishnan, R. and Tickoo, O. Improving model calibration with accuracy versus uncertainty optimization. Advances in Neural Information Processing Systems, 2020.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems, 2017.

Liu, B., Ben Ayed, I., Galdran, A., and Dolz, J. The devil is in the margin: Margin-based label smoothing for network calibration. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. Advances in Neural Information Processing Systems, 2018.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. A simple baseline for Bayesian uncertainty in deep learning. In Advances in Neural Information Processing Systems, 2019.

Musgrave, K., Belongie, S., and Lim, S.-N. Benchmarking validation methods for unsupervised domain adaptation. arXiv preprint arXiv:2208.07360, 2(6):12, 2022.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems, 2019.

Park, S., Bastani, O., Weimer, J., and Lee, I. Calibrated prediction with covariate shift via unsupervised domain adaptation. In International Conference on Artificial Intelligence and Statistics, 2020.

Park, S., Dobriban, E., Lee, I., and Bastani, O. PAC prediction sets under covariate shift. In International Conference on Learning Representations, 2022.

Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., and Saenko, K. Visda: The visual domain adaptation challenge. arXiv preprint arXiv:1710.06924, 2017.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In IEEE/CVF International Conference on Computer Vision, 2019.

Popoviciu, T. Sur certaines inégalités qui caractérisent les fonctions convexes. Analele Stiinti ce Univ. "Al. I. Cuza", Iasi, Sectia Mat, 11:155–164, 1965.

Rahaman, R. et al. Uncertainty quantification and deep ensembles. In Advances in Neural Information Processing Systems, 2021.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet classifiers generalize to ImageNet? In International Conference on Machine Learning, 2019.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115: 211–252, 2015.

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. Maximum classifier discrepancy for unsupervised domain adaptation. In IEEE Conference on Computer Vision and Pattern Recognition, 2018.

Saito, K., Kim, D., Teterwak, P., Sclaroff, S., Darrell, T., and Saenko, K. Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density. In IEEE/CVF International Conference on Computer Vision, 2021.

Salvador, T., Voleti, V., Iannantuono, A., and Oberman, A. Frustratingly easy uncertainty estimation for distribution shift. arXiv preprint arXiv:2106.03762, 2021.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. Advances in Neural Information Processing Systems, 2018.

Shi, Y. and Sha, F. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In International Coference on Machine Learning, 2012.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., et al. Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203, 2019.

Stone, M. Asymptotics for and against cross-validation. Biometrika, pp. 29–35, 1977.

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. Journal of Machine Learning Research, 8(5), 2007.

Thulin, M. The cost of using exact confidence intervals for a binomial proportion. Electronic Journal of Statistics, 8: 817–840, 2014.

Tomani, C., Cremers, D., and Buettner, F. Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. In European Conference on Computer Vision, 2022.

Tu, W., Deng, W., Gedeon, T., and Zheng, L. Assessing model out-of-distribution generalization with softmax prediction probability baselines and a correlation method, 2023. URL https://openreview.net/forum?id=1maXoEyeqx .

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In IEEE Conference on Computer Vision and Pattern Recognition, 2017.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17:261–272, 2020.

Wang, X., Long, M., Wang, J., and Jordan, M. Transferable calibration with lower bias and variance in domain adaptation. In Advances in Neural Information Processing Systems, 2020.

Yang, J., Qian, H., Xu, Y., Wang, K., and Xie, L. Can we evaluate domain adaptation models without target-domain labels? In International Conference on Learning Representations, 2024.

You, K., Wang, X., Long, M., and Jordan, M. Towards accurate model selection in deep unsupervised domain adaptation. In International Conference on Machine Learning, 2019.

Yu, Y., Bates, S., Ma, Y., and Jordan, M. Robust calibration with multi-domain temperature scaling. In Advances in Neural Information Processing Systems, 2022.

Zhang, Y., Liu, T., Long, M., and Jordan, M. Bridging theory and algorithm for domain adaptation. In International Conference on Machine Learning, 2019.

A. Proof of Claims

A.1. Proof of Proposition 2.1

Proof. The proof consists of three parts: 1) decomposition of the expected mean-square error of an estimator $g(x_i; D)$, 2) deriving MLEs of individual and group accuracies; 3) constructing a suf cient condition.

1) Bias-variance decomposition of the expected mean-square error. The expected mean-square error of an estimator $g(x_i; D)$ for (x_i) at $x_i \in G_n^{(t)}$ with respect to the realization of a label $Y|x_i$ can be decomposed by

$$E_D[(\hat{}(x_i) - g(x_i; D))^2] = Var_D(g(x_i; D)) + (Bias_D(g(x_i; D)))^2 + \sigma^2_{x_i} \quad (17)$$

where $Var_D(g(x_i; D)) := E_D[(g(x_i; D) - E_D[g(x_i; D)])^2]$ is the variance of the estimator and $Bias_D(g(x_i; D)) := E_D[g(x_i; D)] - (x_i)$ is the bias of the estimator.

2) MLEs of individual and group accuracy estimators. For an individual accuracy estimator $\hat{}^{(id)}(x; D)$ that predicts an accuracy for each sample x given D, an MLE estimator is $\hat{}^{(id)}(x_i) = \hat{}(x_i)$. This estimator is unbiased because $E_D(\hat{}^{(id)}(x; D)) = (x)$ for each $x \in G_n^{(t)}$. Therefore, this estimator has the average of expected errors

$$\frac{1}{N_n}\sum_{k \in [N_n]} E_D[(\hat{}(x_k) - \hat{}^{(id)}(x_k; D))^2] = \bar{\sigma}^2 + \bar{\sigma}^2 \quad (18)$$

where $\bar{\sigma}^2 := \frac{1}{N_n}\sum_{i \in [N_n]} \sigma^2_{x_i}$.

For a group accuracy estimator $\hat{}^{(gr)}(x; D)$ that predicts the same group accuracy estimate for all $x \in G_n^{(t)}$, an MLE estimator can be de ned by $\hat{}^{(gr)}(x; D) = \frac{1}{N_n}\sum_{i=1}^{N_n} \hat{}(x_i)$, which is a biased estimator because $E_D(\hat{}^{(gr)}(x; D)) = \frac{1}{N_n}\sum_{i=1}^{N_n} (x_i)$ for each $x \in G_n^{(t)}$. Therefore, this estimator has the average of expected errors

$$\frac{1}{N_n}\sum_{k \in [N_n]} E_D[(\hat{}(x_k) - \hat{}^{(gr)}(x_k))^2] = \frac{1}{N_n}\bar{\sigma}^2 + \frac{1}{N_n}\sum_{k \in [N_n]}\left(\frac{1}{N_n}\sum_{i \in [N_n]} (x_i) - (x_k)\right)^2 + \bar{\sigma}^2 \quad (19)$$

$$= \frac{1}{N_n}\bar{\sigma}^2 + Var(\ ; D) + \bar{\sigma}^2 \quad (20)$$

where $Var(\ ; D)$ is the variance of the accuracy in group $G_n^{(t)}$.

3) Suf cient condition. Given (18) and (20), the Popoviciu's inequality (Popoviciu, 1965) provides a suf cient condition for the group accuracy estimator $\hat{}^{(gr)}$ to have a lower expected mean-squared error than the individual accuracy estimator $\hat{}^{(id)}$ as follows:

$$Var(\ ; D) \quad \frac{1}{4}\left(\max_{x' \in G_n^{(t)}} (x') - \min_{x' \in G_n^{(t)}} (x')\right)^2 \quad \frac{N_n - 1}{N_n}\bar{\sigma}^2 = \frac{N_n - 1}{N_n}\left(\frac{1}{N_n}\sum_{i \in [N_n]} (x_i)(1 - (x_i))\right) \quad (21)$$

where the equality comes from $\bar{\sigma}^2 = \frac{1}{N_n}\sum_{i \in [N_n]} \sigma^2_{x_i}$ with $\sigma^2_{x_i}$ is the variance of the Bernoulli distribution with a parameter (x_i). □

A.2. Suf cient Condition for (1)

To nd the suf cient condition for (1), i.e., $\frac{1}{4}(\max_{x' \in G_n} (x') - \min_{x' \in G_n} (x'))^2 \quad \frac{N_n - 1}{N_n}\bar{\sigma}^2$, note that $\bar{\sigma}^2 \min_{x \in G_n} \sigma^2_x = \tilde{}(1 - \tilde{})$ where $\tilde{} = \min\{\min_{x' \in G_n} (x'); (1 - \max_{x' \in G_n} (x'))\}$. Therefore, the suf cient condition is $\max_{x' \in G_n} (x') - \min_{x' \in G_n} (x') \quad 2\sqrt{\frac{N_n - 1}{N_n}\tilde{}(1 - \tilde{})}$, which is described in Figure A1.

A.3. Decomposition of the Expected Squared Calibration Error

Proposition A.1. The expected squared calibration error can be decomposed as the sum of the variance of the accuracies for samples in the same group and a squared group accuracy estimation error. That is,

$$E_{p_T}[(P(Y = \hat{Y}) - \hat{}_T(G^{(t)}_{l_{(g)}(X)}))^2] = \sum_{n \in [M]} P(X_T \in G_n^{(t)})[Var(P(Y = \hat{Y})|G_n^{(t)}) + (\ _T(G_n^{(t)}) - \hat{}_T(G_n^{(t)}))^2] \quad (22)$$

Figure A1. The shaded area includes values of maximum and minimum expected accuracies within the group, which satisfy the sufficient condition for (1) when $N_n = 100$.

where X_T is a random variable having a density p_T.

Proof.

$$E_{p_T}[(P(Y = \hat{Y}) - \hat{\alpha}_T(G^{(t)}_{l_{(g)}(X)}))^2] = \sum_{n \in [M]} P(X_T \in G^{(t)}_n) E_{p_T}[(P(Y = \hat{Y}) - \hat{\alpha}_T(G^{(t)}_{l_{(g)}(X)}))^2 | X \in G^{(t)}_n] \quad (23)$$

$$= \sum_{n \in [M]} P(X_T \in G^{(t)}_n) E_{p_T}[(P(Y = \hat{Y}) - \alpha_T(G^{(t)}_n) + \alpha_T(G^{(t)}_n) - \hat{\alpha}_T(G^{(t)}_n))^2 | X \in G^{(t)}_n] \quad (24)$$

$$= \sum_{n \in [M]} P(X_T \in G^{(t)}_n) \left(E_{p_T}[(P(Y = \hat{Y}) - \alpha_T(G^{(t)}_n))^2 | X \in G^{(t)}_n] + (\alpha_T(G^{(t)}_n) - \hat{\alpha}_T(G^{(t)}_n))^2 \right) \quad (25)$$

$$= \sum_{n \in [M]} P(X_T \in G^{(t)}_n)[Var(P(Y = \hat{Y}) | G^{(t)}_n) + (\alpha_T(G^{(t)}_n) - \hat{\alpha}_T(G^{(t)}_n))^2] \quad (26)$$

where the equality (25) holds due to $E_{p_T}[(P(Y(X) = \hat{Y}(X)) - \alpha_T(G^{(t)}_n))(\alpha_T(G^{(t)}_n) - \hat{\alpha}_T(G^{(t)}_n)) | X \in G^{(t)}_n] = 0$.

□

A.4. Proof of Proposition 3.1

Proof. By applying triangle inequalities, we get the following inequality:

$$\begin{aligned} |\alpha_S(G^{(t)}_n; w) - \alpha_S(G^{(t)}_n; w^y(n))| \le{}& |\alpha_S(G^{(t)}_n; w) - \hat{\alpha}^{(MC)}_S(G^{(t)}_n)| + |\hat{\alpha}^{(MC)}_S(G^{(t)}_n) - \hat{\alpha}^{(IW)}_S(G^{(t)}_n; w^y(n))| \\ &+ |\hat{\alpha}^{(IW)}_S(G^{(t)}_n; w^y(n)) - \alpha^{(IW)}_S(G^{(t)}_n; w^y(n))| + |\alpha^{(IW)}_S(G^{(t)}_n; w^y(n)) - \alpha_S(G^{(t)}_n; w^y(n))|. \end{aligned} \quad (27)$$

Note that the first and third terms in the right hand side are coming from the Monte-Carlo approximation, so they can be bounded by $O(\log(1/\tilde{\delta})/|G^{(t)}_n(D_S)|)$ with probability at least $1 - \tilde{\delta}$ based on a concentration inequality such as the Hoeffding's inequality. Also, the second term is bounded by the optimization error $\mathrm{err}_{\mathrm{opt}}(w^y(n))$. Therefore, it is enough to analyze the fourth term.

The fourth term is coming from the bias of $E_{T_{Y|X}}[1(Y(X) = \hat{Y}(X))] = \alpha_T(G^{(t)}_n; w^y(n))$, which we refer to as the bias of the identical accuracy assumption. It can be bounded by

$$|\alpha^{(IW)}_S(G^{(t)}_n; w^y(n)) - \alpha_S(G^{(t)}_n; w^y(n))| \quad (28)$$

$$= \frac{P(X_T \in G^{(t)}_n)}{P(X_S \in G^{(t)}_n)} E_{p_T}\left[\frac{1(Y(X) = \hat{Y}(X)) - \alpha_T(G^{(t)}_n; w^y(n))}{w^y(n)(X)} \middle| G^{(t)}_n \right] \quad (29)$$

$$\frac{P(X_T \in G_n^{(t)})}{P(X_S \in G_n^{(t)})} \left[E_{p_T} \left[(1(Y(X) = \hat{Y}(X)) - \hat{\alpha}_T(G_n^{(t)}; w^y(n)))^2 \mid G_n^{(t)} \right] E_{p_T} \left[\frac{1}{w^y(n)(X)^2} \mid G_n^{(t)} \right] \right]^{1/2} \tag{30}$$

$$\frac{P(X_T \in G_n^{(t)})}{2P(X_S \in G_n^{(t)})} \left[E_{p_T} \left[(1(Y(X) = \hat{Y}(X)) - \hat{\alpha}_T(G_n^{(t)}; w^y(n)))^2 \mid G_n^{(t)} \right] + E_{p_T} \left[\frac{1}{w^y(n)(X)^2} \mid G_n^{(t)} \right] \right] \tag{31}$$

$$\frac{P(X_T \in G_n^{(t)})}{2P(X_S \in G_n^{(t)})} \left[E_{p_T} \left[(1(Y = \hat{Y}) - \hat{\alpha}_T(G_n^{(t)}; w^y(n)))^2 \mid G_n^{(t)} \right] + \frac{1}{w^y(n)^2} \right] \tag{32}$$

where (30) holds due to the Cauchy-Schwarz inequality, (31) holds due to the AM-GM inequality, and $w^y(n) := \min_{i \in [2B]} \{ w_i^y(n) \}$. □

A.5. Formal Statement and Proof of Proposition A.2

Proposition A.2 (Bias-variance decomposition). Let $\hat{\alpha}_T(G_n^{(t)})$ be an estimate for $\alpha_T(G_n^{(t)}; w)$. Then, the bias of the identical accuracy assumption is given by

$$\mathrm{IdentBias}(w^y(n); G_n^{(t)}) = \frac{P(X_T \in G_n^{(t)})}{2P(X_S \in G_n^{(t)})} \left[\frac{1}{w^y(n)^2} + \mathrm{Bias}(\hat{\alpha}_T(G_n^{(t)}))^2 + \mathrm{Var}(1(Y = \hat{Y}) \mid G_n^{(t)}) \right] \tag{33}$$

where $\mathrm{Bias}(\hat{\alpha}_T(G_n^{(t)})) := | \alpha_T(G_n^{(t)}; w) - \hat{\alpha}_T(G_n^{(t)}) |$ is the bias of the estimate $\hat{\alpha}_T(G_n^{(t)})$ and $\mathrm{Var}(1(Y = \hat{Y}) \mid G_n^{(t)}) := E_{p_T} \left[1(Y = \hat{Y}) - \alpha_T(G_n^{(t)}; w) \right]^2 \mid G_n^{(t)} \right]$ is the variance of the correctness of predictions on $G_n^{(t)}$.

Proof. Based on the proof of Proposition 3.1, it is enough to decompose $E_{p_T}[(1(Y(X) = \hat{Y}(X)) - \hat{\alpha}_T(G_n^{(t)}))^2]$ as follows

$$E_{p_T} \left[(1(Y(X) = \hat{Y}(X)) - \hat{\alpha}_T(G_n^{(t)}))^2 \mid G_n^{(t)} \right] \tag{34}$$

$$= E_{p_T} \left[(1(Y(X) = \hat{Y}(X)) - \alpha_T(G_n^{(t)}; w) + \alpha_T(G_n^{(t)}; w) - \hat{\alpha}_T(G_n^{(t)}))^2 \mid G_n^{(t)} \right] \tag{35}$$

$$= E_{p_T} \left[(1(Y(X) = \hat{Y}(X)) - \alpha_T(G_n^{(t)}; w))^2 + (\alpha_T(G_n^{(t)}; w) - \hat{\alpha}_T(G_n^{(t)}))^2 \mid G_n^{(t)} \right] \tag{36}$$

where the equality (36) holds due to $E_{p_{T_X}} E_{p_{T_{Y|X}}} [(1(Y(X) = \hat{Y}(X)) - \alpha_T(G_n^{(t)}; w))(\alpha_T(G_n^{(t)}; w) - \hat{\alpha}_T(G_n^{(t)}))] \mid G_n^{(t)}] = 0$. □

B. Discussions

B.1. Model Calibration in the I.I.D. Settings

In a classi cation problem, the maximum value of the softmax output is often considered as a con dence of a neural network's prediction. In (Guo et al., 2017), it is shown that the modern neural networks are poorly calibrated, tending to produce larger con dences than their accuracies. Based on this observation, (Guo et al., 2017) introduce a post-processing approach that adjusts a temperature parameter of the softmax function for adjusting the overall con dence level. In Bayesian approaches (such as Monte-Carlo dropout (Gal & Ghahramani, 2016; Gal et al., 2017), deep ensemble (Lakshminarayanan et al., 2017; Rahaman et al., 2021), and a last-layer Bayesian approach (Sensoy et al., 2018; Joo et al., 2020)), the con dence level adjustment is induced by posterior inference and model averaging. While both post-hoc calibration methods and Bayesian methods have been achieving impressive calibration performances in the i.i.d. setting (Maddox et al., 2019; Ovadia et al., 2019; Ebrahimi et al., 2020), it has been shown that most of the calibration improvement methods fall short under distribution shifts (Ovadia et al., 2019).

B.2. On Choice of Non-Parametric Estimators

Our concept of determining the IW from its CI can be applied to any other valid CI estimators. For example, by analyzing a CI of the odds ratio of the logistic regression used as a domain classi er (Bickel et al., 2007; Park et al., 2020; Salvador et al., 2021), a CI of the IW can be obtained. Then, IW-GAE can be applied in the same way as developed in Section 3. As an extreme example, we apply IW-GAE by setting minimum and maximum values of IWs as CIs in an ablation study (Table

4). While IW-GAE outperforms strong baseline methods (CPCS and TransCal) even under this naive CI estimation, we observe that its performance is reduced compared to the setting with a sophisticated CI estimation (Park et al., 2022). In this regard, advancements in IW estimation or CI estimation would be bene cial for accurately estimating the group accuracy, thereby model selection and uncertainty estimation. Therefore, we leave combining IW-GAE with advanced IW estimation techniques as an important future direction of research.

B.3. On Choice of the Number of Groups

In this work, we estimate the group accuracy by grouping predictions based on the con dence of the prediction. Therefore, a natural question to ask is how to select the number of groups. If we use a small number of groups, then there would be high $\mathrm{IdentBias}(w^y; G_n^{(t)})$ because of the large variance of prediction correctness within a group. In addition, reporting the same accuracy estimate for a large number of predictions could be inaccurate in terms of representing uncertainty for individual predictions. Conversely, if we use a large number of bins, there would be high Monte-Carlo approximation $\varepsilon_{\text{stat}}$ errors, Therefore, it would result in a loose connection between the source group accuracy estimation error and the objective in the optimization problem (cf. Proposition 3.1). Based on our experimental results with $M = 10$ and the sensitivity analysis (cf. Section 4.4), we would recommend to choose M from its standard range $[10, 20]$ in the literature.

C. Additional Details

C.1. Obtaining CIs of Binned IWs

In this work, we use a recently proposed nonparametric estimation method (Park et al., 2022) for constructing the CI of the IW[1]. In this approach, $\mathcal{X}$ is partitioned into B number of bins $\mathcal{X} = [_{i=1}^{B} B_i)$ where $B_i = \{x \in \mathcal{X} \mid I^{(B)}(x) = i\}$. Here, the partition function is de ned such that $I^{(B)}(x) = j$ if $q(\frac{j-1}{B}; W) < w(x) \leq q(\frac{j}{B}; W)$ where $W := \{w(x) \mid x \in D_S \cup D_T\}$ and w is a rough estimate of IW (cf. Appendix C.2).

Then, con dence intervals (CIs) of the binned probabilities $p_S(x) = p_{S_{I^{(B)}(x)}}$ with $p_{S_j} = \int_{B_j} p_S(x)dx$ and $p_T(x) = p_{T_{I^{(B)}(x)}}$ with $p_{T_j} = \int_{B_j} p_T(x)dx$ are constructed. Speci cally, for p_{S_j}, the number of samples in a bin $n_j^{(S)} := \sum_{i=1}^{N^{(S)}} \mathbb{1}(x_i^{(S)} \in B_j)$ is interpreted as a sample from $\mathrm{Binom}(N^{(S)}; p_{S_j})$. Then, the Clopper–Pearson CI (Clopper & Pearson, 1934) provides the CI of p_{S_j} as $[\ell_S(n_j^{(S)}; N^{(S)}, \delta/2), u_S(n_j^{(S)}; N^{(S)}, \delta/2)]$ with probability at least $1-\delta$ where $\ell(k; m, \alpha) := \inf\{\theta \in [0,1] \mid F(k; m, \theta) \leq \alpha\}$ and $u(k; m, \alpha) := \sup\{\theta \in [0,1] \mid F(k; m, \theta) \geq \alpha\}$ with F being the cumulative distribution function of the binomial distribution. Similarly, we can obtain the CIs of p_{T_j} by collecting $n_j^{(T)} := \sum_{i \in [N^{(T)}]} \mathbb{1}(x_i^{(T)} \in B_j)$ and following the same procedure, which are denoted as $[\ell_T(n_j^{(T)}; N^{(T)}, \delta/2), u_T(n_j^{(T)}; N^{(T)}, \delta/2)]$.

With the CIs of p_{S_j} and p_{T_j} for $j \in [B]$, the CI of the IW in B_j can be obtained. Speci cally, for $\delta := \delta/2B$, the following inequality holds with probability at least $1-\delta$ (Park et al., 2022):

$$w_j := \frac{p_{T_j}}{p_{S_j}} \in \left[\frac{[\ell_T(n_j^{(T)}; N^{(T)}, \delta) - G]^+}{u_S(n_j^{(S)}; N^{(S)}, \delta) + G}, \frac{u_T(n_j^{(T)}; N^{(T)}, \delta) + G}{[\ell_S(n_j^{(S)}; N^{(S)}, \delta) - G]^+}\right] \tag{37}$$

where $[a]^+ := \max\{0, a\}$ and $G \in \mathbb{R}_+$ is a constant that satis es $\int_{B_j} |p_S(x) - p_S(x')|dx' \leq G$ and $\int_{B_j} |p_T(x) - p_T(x')|dx' \leq G$ for all $x \in B_j$ and $j \in [B]$. We refer to $\{w_i\}_{i \in B}$ as binned IWs. Also, we de ne the CI of w_j as

$$\mathcal{I}_j := \left[\frac{[\ell_T(n_j^{(T)}; N^{(T)}, \delta) - G]^+}{u_S(n_j^{(S)}; N^{(S)}, \delta) + G}; \frac{u_T(n_j^{(T)}; N^{(T)}, \delta) + G}{[\ell_S(n_j^{(S)}; N^{(S)}, \delta) - G]^+}\right]. \tag{38}$$

C.2. IW Estimator

IW estimation is required for implementing baseline methods and construct bins for estimating the CI of the IW. We adopt a logistic regression model on top of the neural network's representation as the discriminative learning-based estimation

[1]We note that IW-GAE works with any interval estimation methods (cf. Discussions in Appendix B.2) or even arbitrary intervals (cf. Section 4.3).

(Bickel et al., 2007), following Wang et al. (2020). Specifically, it first upsamples from one domain to match the sizes, i.e., $|D_S| = |D_T|$, and then it labels samples with the domain index, i.e., $\{(h(x), 1) | x \in D_T\}$ and $\{(h(x), 0) | x \in D_S\}$ where h is the feature map of the neural network. After training the logistic regression model v with a quasi-Newton method until convergence, the IW can be estimated as $\hat{w}(x) = v(h(x))/(1 - v(h(x)))$.

C.3. Algorithm

Algorithm 1 Pseudocode of IW-GAE

Input: Source dataset $D_S = \{(x_i^{(S)}, y_i^{(S)})\}_{i=1}^{N^{(S)}}$, Target dataset $D_T = \{x_i^{(T)}\}_{i \in [N^{(T)}]}$
Hyperparameters: The numbers of bins and groups B (and M), level of CI α, search space $\mathcal{T}$
\# Prepare a UDA model (Wang et al., 2020)
Partition D_S into D_S^{tr} and D_S^{val}
Train a neural network g on (D_S^{tr}, D_T) with any UDA method
Upsample D_S^{tr} or D_T to make $|D_S^{tr}| = |D_T|$
Compute $F_S^{tr} = \{g(x) | x \in D_S^{tr}\}$, $F_S^{val} = \{g(x) | x \in D_S^{val}\}$, and $F_T = \{g(x) | x \in D_T\}$
Train a logistic regression model h that discriminates F_S^{tr} and F_T
\# Obtain CIs of binned IWs (Park et al., 2022)
Gather IWs $W_{S[T]} = \{(1 - H(g(x)))/H(g(x)) : x \in D_S^{val}[D_T]\}$
Compute quantiles $q(i) = i/(B+1)$-th quantile of $W_{S[T]}$ for $i \in [B+1]$
Construct bins $B_i = \{x \in D_S^{val}[D_T : q(i) \le (1 - H(g(x)))/H(g(x)) \le q(i+1)\}$ for $i \in [B]$
Compute δ_i using (37) for each $i \in [B]$
\# IW-GAE
$\epsilon^y = 1$
for $t \in \mathcal{T}$ **do**
  Obtain $w(n;t)$ by solving the optimization problem in (8)-(13) for $n \in [M]$
  if $\sum_{n \in [M]} (\hat{\alpha}_S^{(MC)}(G_n^{(t)}) - \hat{\alpha}_S^{(IW)}(G_n^{(t)}; w(n;t)))^2 < \epsilon^y$ **then**
    $\epsilon^y = \sum_{n \in [M]} (\hat{\alpha}_S^{(MC)}(G_n^{(t)}) - \hat{\alpha}_S^{(IW)}(G_n^{(t)}; w(n;t)))^2$
    $t^y = t$
    $w^y(n;t^y) = w(n;t)$ for $n \in [M]$
  end if
end for

D. Experimental Details

We follow the exact same training configurations as those used in the Transfer Learning Library (Jiang et al., 2020), except we separate 20% as the validation dataset from the source domain (in the original implementation, validation is performed with the test dataset for OfficeHome). The configuration of training MDD for OfficeHome is as follows: MDD is trained for 30 epochs with SGD with momentum parameter 0.9 and weight decay 0.0005. The learning rate is schedule by $\eta_0(1 + \gamma t)^{-\rho}$ where t is the iteration counter, $\eta_0 = 0.004$, $\gamma = 0.0002$, $\rho = 0.75$, and the stochastic gradient is computed with minibatch of 32 samples from the source domain and 32 samples from the target domain. Also, it uses the margin coefficient of 4 as the MDD-specific hyperparamter. For the model architecture, it uses ResNet-50 pre-trained on ImageNet (Russakovsky et al., 2015) with the bottleneck dimension of 2,048. For the large-scale VisDa-2017, only the architecture is changed to ResNet-101 with the bottleneck dimension of 1,024 under the same training configuration.

IW-GAE specific details For CI estimation, we follow the same configuration with the original method (Park et al., 2022). Specifically, we use constant $\delta = 0.001$, CI level $\alpha = 0.05$, and the number of bins $B = 10$. In addition, we use the maximum IW value $w_n^{(ub)} = 6.0$ and the minimum IW value $w^y(n) = 1/6$ for $n \in [M]$, which is a common technique in IW-based estimations (Wang et al., 2020; Park et al., 2022). For the constraint relaxation constants in IW-GAE, we use $\epsilon_{tol} = 0.1$ and $\epsilon_{pr} = 0.3$. For implementing sequential least square programming, we use the SciPy Library (Virtanen et al., 2020) with tolerance 10^{-8} that is used to check a convergence condition (other optimizer-specific values follow the default values in SciPy) and choose the middle points from CIs of binned IW as an initial solution.

E. Missing Tables

Table A1.Model calibration benchmark results of MDD with ResNet-50 on Of ce-Home. We repeat experiments for ten times and report the average value of ECE.

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
Vanilla	40.61	25.62	15.56	33.83	25.34	24.75	33.45	38.62	16.76	23.37	36.51	14.13	27.37
TS	35.86	22.84	10.60	28.24	20.74	20.06	32.47	37.20	14.89	18.36	34.62	12.28	24.01
CPCS	22.93	22.07	10.19	26.88	18.36	14.05	28.28	29.20	12.06	15.76	26.54	11.19	19.79
IW-TS	32.63	22.90	11.27	28.05	19.65	18.67	30.77	38.46	15.10	17.69	32.20	12.37	23.76
TransCal	33.57	20.27	8.88	26.36	18.81	18.42	27.35	29.86	10.48	16.17	29.90	10.08	20.84
PTS	31.91	24.36	10.65	22.81	20.42	15.92	26.55	39.34	12.38	21.83	26.31	10.14	21.91
AvUTS	29.59	25.55	10.40	31.81	26.06	26.15	37.10	46.04	20.75	27.70	41.27	15.61	28.17
IW-Mid	23.25	31.62	12.99	17.15	18.71	9.23	27.75	30.35	9.02	13.64	26.32	10.19	19.22
IW-GAE	12.78	4.70	12.93	7.52	4.42	4.11	9.50	17.49	8.40	7.62	9.52	8.14	8.93
Oracle	10.45	10.72	6.47	8.10	7.62	6.55	11.88	9.39	5.93	7.54	10.72	5.76	8.42

Table A2.Large-scale model calibration benchmark results of CDAN with ResNet-50 on DomainNet. The numbers indicate the mean ECE across ten repetitions. Cl, Pt, Rw, and Sk correspond to clipart, painting, real, and sketch, respectively.

Method	Cl-Pt	Cl-Rw	Cl-Sk	Pt-Cl	Pt-Rw	Pt-Sk	Rw-Cl	Rw-Pt	Rw-Sk	Sk-Cl	Sk-Pt	Sk-Rw	Avg
Vanilla	13.23	6.36	12.92	9.75	6.35	15.56	9.44	9.70	14.34	6.63	11.25	5.12	10.06
TS	12.95	5.95	13.32	6.40	3.90	11.07	8.64	10.49	16.08	3.17	5.58	13.09	9.22
CPCS	5.64	21.90	7.70	5.14	7.72	7.90	9.35	11.17	17.06	3.46	2.23	15.90	9.60
IW-TS	16.76	16.70	12.53	5.29	7.84	4.34	9.60	10.58	16.80	5.40	2.98	17.10	10.49
TransCal	18.51	29.63	20.92	23.02	31.83	17.58	27.88	28.83	20.31	31.66	23.06	32.56	25.39
IW-Mid	7.61	11.01	5.89	8.84	7.58	5.36	8.70	7.49	7.53	10.24	8.10	10.28	8.21
IW-GAE	6.06	8.15	5.38	7.45	3.89	3.94	7.01	5.58	6.73	6.80	6.82	8.06	6.32
Oracle	4.55	2.78	4.01	3.10	3.72	2.72	3.10	2.79	2.83	3.13	1.70	1.73	3.02

Table A3.Hyperparameter selection benchmark results of MDD with ResNet-50 on Of ce-Home. We repeat experiments for ten times and report the average test accuracy of selected model.

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
Vanilla	53.31	70.96	77.44	59.70	65.17	69.96	57.07	50.95	74.75	68.81	57.11	80.13	65.45
IWCV	53.24	69.61	72.50	59.70	65.17	67.50	57.07	55.21	74.75	68.81	58.51	80.13	65.18
DEV	53.31	70.72	77.44	59.79	67.99	69.96	57.07	52.50	77.12	70.50	53.38	82.27	66.70
InfoMax	54.34	70.96	77.53	61.48	69.93	71.06	62.79	54.41	78.79	71.32	58.51	82.37	67.79
SND	44.55	68.14	75.57	58.86	66.04	66.46	61.13	53.20	70.38	62.54	56.15	80.26	63.60
Corr-C	50.88	70.96	78.47	60.74	68.60	71.06	62.79	54.41	78.79	71.32	58.51	82.64	67.41
TransScore	54.34	70.96	77.53	61.48	69.93	71.06	62.79	54.41	78.79	71.32	58.51	82.36	67.87
MixVal	54.34	70.09	77.43	60.74	59.34	70.17	61.00	55.21	78.35	71.32	57.89	82.65	67.05
IW-Mid	54.13	69.27	78.47	61.48	68.03	71.06	59.99	55.21	78.79	70.50	57.11	83.12	67.26
IW-GAE	54.34	70.96	78.47	61.48	69.93	71.06	62.79	55.21	78.79	70.50	58.51	83.39	67.95
Lower bound	52.51	69.27	72.50	59.70	65.17	67.50	57.07	50.95	74.75	68.81	50.90	80.13	64.10
Oracle	54.34	70.96	78.47	61.48	69.93	71.06	62.79	55.21	78.79	71.32	58.51	83.01	67.91

Table A4.Checkpoint selection benchmark results of MDD with ResNet-50 on Of ce-Home. We repeat experiments for ten times and report the average test accuracy of selected model.

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
Vanilla	47.22	74.14	77.76	61.85	70.96	71.59	60.98	53.63	78.93	71.57	57.04	83.97	67.47
IWCV	54.46	74.22	72.27	61.48	70.49	70.62	61.30	51.13	78.37	72.94	58.43	84.07	67.48
DEV	54.04	73.94	78.16	61.52	63.19	70.70	60.43	53.63	78.93	71.57	58.62	83.97	67.39
InfoMax	54.32	74.72	77.90	62.79	71.03	71.47	61.39	53.15	78.75	72.89	58.53	83.98	67.38
SND	47.67	73.06	77.71	62.67	70.47	71.17	61.43	52.14	78.75	70.71	57.66	79.97	67.23
Corr-C	54.46	74.72	77.53	61.76	70.88	71.24	61.30	52.47	78.40	72.59	58.53	83.89	67.24
TransScore	54.79	74.14	77.77	61.76	70.97	71.48	61.17	53.15	78.93	72.89	58.53	83.98	67.38
MixVal	54.45	74.35	77.77	61.93	70.70	71.43	61.30	53.29	78.93	72.89	58.53	83.89	67.37
IW-Mid	54.04	72.63	78.37	62.05	71.28	71.45	61.25	54.39	79.07	73.19	58.75	80.63	67.04
IW-GAE	54.32	73.98	78.51	61.96	71.25	71.70	61.10	54.30	78.91	73.22	58.70	83.68	67.48
Lower bound	41.90	64.88	72.27	52.00	58.48	62.13	53.52	38.33	70.92	63.41	44.81	75.58	58.21
Oracle	54.80	74.79	78.61	62.79	71.59	72.18	61.64	54.64	79.44	73.42	59.43	84.62	68.25

Table A5.Model calibration benchmark results of CDAN with ResNet-50 on Of ce-Home. The numbers indicate the mean ECE across ten repetitions.

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
Vanilla	30.73	18.38	14.37	25.63	22.44	19.10	27.54	36.72	12.48	19.93	31.12	10.28	22.84
TS	29.68	19.40	14.40	22.15	19.97	16.88	28.82	38.03	12.99	20.46	31.91	11.28	22.21
CPCS	18.78	18.09	14.74	22.18	20.74	16.33	29.30	34.92	11.92	20.99	31.41	11.07	21.87
IW-TS	12.38	16.79	14.85	21.75	20.06	16.92	29.30	38.84	13.30	20.82	31.10	12.36	22.62
TransCal	7.94	14.05	12.91	7.82	9.25	10.23	9.37	12.60	14.29	9.92	9.76	1.51	10.30
IW-Mid	36.05	47.70	26.82	21.08	22.95	21.55	18.88	28.99	15.39	21.16	28.16	2.67	25.17
IW-GAE	13.98	29.82	9.44	6.55	5.59	10.16	5.29	13.47	11.01	11.12	7.26	1.41	9.13
Oracle	7.91	8.80	6.05	7.57	7.93	6.76	9.07	9.14	4.04	7.16	9.19	5.67	7.44

Table A6.Model calibration benchmark results of MCD with ResNet-50 on Of ce-Home. The numbers indicate the mean ECE across ten repetitions.

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
Vanilla	38.91	26.39	18.86	32.85	26.69	19.36	35.87	36.70	18.61	24.57	36.87	14.79	27.54
TS	31.84	22.55	13.49	26.16	20.10	10.72	33.98	31.91	15.59	21.62	31.59	12.46	22.67
CPCS	13.07	20.09	47.15	9.78	21.82	8.02	32.65	25.61	15.27	20.53	40.38	7.38	23.85
IW-TS	12.88	21.44	61.15	10.56	16.40	11.72	33.03	36.37	14.09	19.96	41.95	12.80	24.91
TransCal	19.23	15.09	6.55	17.91	11.60	3.91	22.98	15.81	6.11	13.77	21.40	4.62	11.20
IW-Mid	50.68	28.93	23.92	38.24	33.48	28.58	39.76	37.45	22.40	27.15	44.15	13.07	32.73
IW-GAE	22.21	10.68	2.38	15.96	9.30	3.53	23.54	22.73	6.37	11.78	20.75	1.63	11.57
Oracle	5.88	9.91	3.19	7.75	4.64	3.66	4.17	7.70	3.09	4.51	8.09	3.54	4.51

F. Additional Experiments

In this section, we show the effectiveness of IW-GAE with two different base models. First, we perform additional experiments with conditional domain adversarial network (CDAN; (Long et al., 2018)) which is also a popular UDA method. As in the experiments with MDD, we use ResNet-50 as the backbone network and Of ceHome as the dataset. The learning rate schedule for CDAN is $(1 + \quad t)$ wheret is the iteration counter, $= 0:01$, $= 0:001$, and $= 0:75$. The remaining training con guration for CDAN is the same as the MDD training con guration except it uses the bottleneck dimension of 256 and weight decay of0.0005(cf. Appendix D). As we can see from Table A5, IW-GAE achieves the best performance among all considered methods, achieving the best ECE in 8 out of the 12 cases as well as the lowest mean ECE. We note that TransCal achieves a performance comparable to IW-GAE in this experiment, but considering the results in the other tasks, IW-GAE is still an appealing method for performing the model calibration task.

We also perform additional experiments with maximum classi er discrepancy (MCD; (Saito et al., 2018)). Following the previous experiments with MDD and CDAN, we use ResNet-50 as the backbone network and Of ceHome as the dataset. The training con guration is the same as the MDD training con guration except it uses the xed learning rate of 0.001 with weight decay of0:0005and bottleneck dimension of 1,024 (cf. Appendix D). Consistent to other benchmark results, IW-GAE achieves the best performance among all methods (Table A6). Speci cally, IW-GAE achieves the best average model calibration performance, and its ECE is lowest in 7 out of 12 domain pairs. Note that IW-Mid's performance with MCD is signi cantly lower compared to other benchmark results. However, IW-GAE still signi cantly improves the performance, indicating that IW-GAE does not strongly depends on accuracy of the CI estimation discussed in Section C.1.

F.1. Qualitative Evaluation of IW-GAE

To qualitatively analyze IW-GAE, we also visualize reliability curves that compare the estimated group accuracy with the average accuracy in Figure A2. We rst note that IW-GAE tends to accurately estimate the true group accuracy for most groups under different cases compared to IW-Mid. The accurate group accuracy estimation behavior of IW-GAE explains the results that the IW-GAE improves IW-Mid for most cases in the model calibration and selection tasks (cf. Tables A1-A6). For most cases, true accuracy is in between the lower and upper IW estimators, albeit the interval length tends to increase for high-con dence groups. This means that the CI of the IW based on the Clopper-Pearson method successfully captures the IW in the CI. We also note that the true accuracy is close to the lower IW estimator in the lower con dence group and the middle IW estimator in the high con dence group. An observation that the true accuracy's relative positions in CIs varies from one group to another group motivates why an adaptive selection of binned IWs as ours is needed.

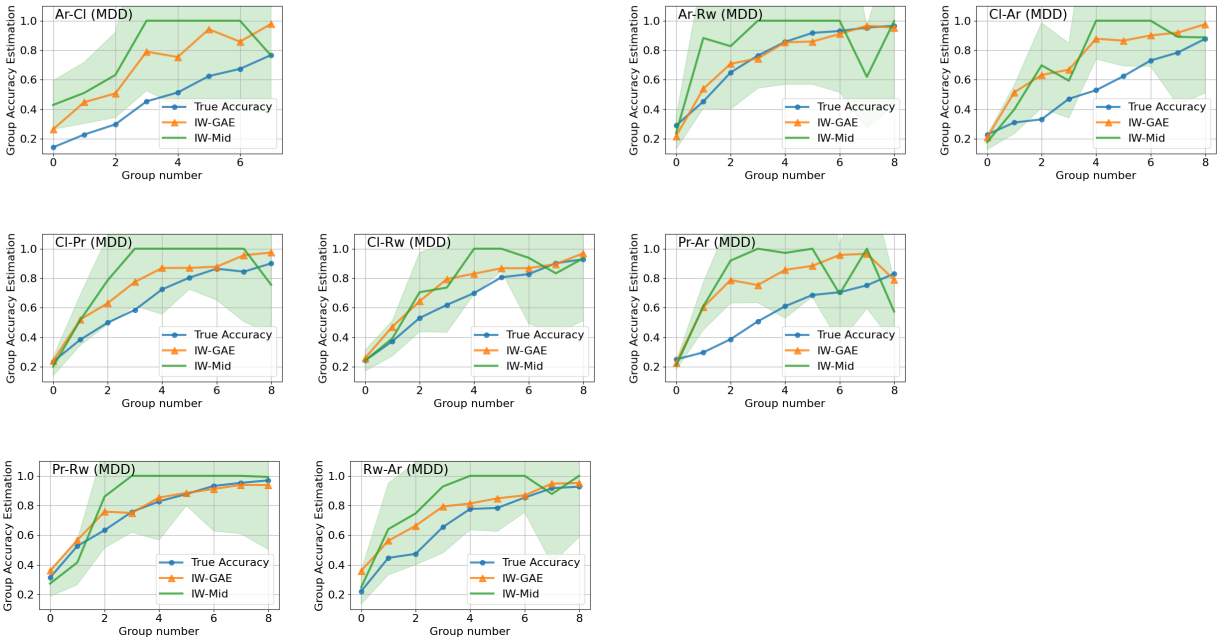


Figure A2. True group accuracy and estimated group accuracy of IW-GAE and IW-Mid under MDD. The shaded areas represent possible group accuracy estimation with binned IWs in the CI. The title of a figure represents "Source-Target." For IW-Mid and IW-GAE, we clip the accuracy estimations when they exceed 1, which can occur when the upper bound of CI is large. Also, the number of groups in the figure is different for some domain pairs because there can be a group that contains no target samples (we set $M = 10$ for all cases).

G. Analysis of $\epsilon_{opt}(w^\dagger(n))$ and $IdentBias(w^\dagger(n); G_n)$ of IW and Their Relation to Source and Target Group Accuracy Estimation Errors

In this section, we aim to answer the following question about the central idea of this work: *"Does solving the optimization problem in (8)-(13) result in an accurate target group accuracy estimator?"* Specifically, we analyze the relationship between the optimization error $_{opt}(w^y(n))$, the bias of the identical accuracy assumption $IdentBias(w^y(n); G_n)$, the source group accuracy estimation error $j \ _S(G_n; w \ ) \quad _S(G_n; w^y(n))j$, and the target group accuracy estimation error $j \ _T(G_n; w \ ) \quad _T(G_n; w^y(n))j$ from the perspective of (5) and (15)[2]. To this end, we gather $w^y(n)$ obtained by solving the optimization problem under all temperature parameters in the search space $t \ 2 \ T$ with MDD on the OfficeHome dataset (720 IWs from 6 values of the temperature parameter, 12 domain pairs, and 10 groups). Then, by using the test dataset in the source and the target domains, we obtain the following observations.

In (5), we show that $j \ _T(G_n; w \ ) \quad _T(G_n; w^y(n))j$ is upper bounded by $j \ _S(G_n; w \ ) \quad _S(G_n; w^y(n))j$. However, the inequality could be loose since the inequality is obtained by taking the maximum over the IW values. Considering that the optimization problem is formulated for finding $w^y(n)$ that achieves small $j \ _S(G_n; w \ ) \quad _S(G_n; w^y(n))j$ (cf. Proposition 3.1), the loose connection between the source and target group accuracy estimation errors can potentially enlighten a fundamental difficulty to our approach. However, as we can see from Figure A4, it turns out that $j \ _S(G_n; w \ )$ $_S(G_n; w^y(n))j$ is strongly correlated with $j \ _T(G_n; w \ ) \quad _T(G_n; w^y(n))j$. This result validates our approach of reducing the source accuracy estimation error of the IW-based estimator for obtaining an accurate group accuracy estimator in the target domain.

In (15), we show that $j \ _S(G_n; w \ ) \quad _S(G_n; w^y(n))j \quad _{opt}(w^y(n)) + \ _{stat} + IdentBias(w^y(n); G_n)$, which motivates us to solve the optimization problem for reducing $_{opt}(w^y(n))$ (cf. Section 3) and to construct groups based on the maximum value of softmax for reducing $IdentBias(w^y(n); G_n)$ (cf. Section 2.2). Again, if these terms are loosely connected to $j \ _S(G_n; w \ ) \quad _S(G_n; w^y(n))j$, a fundamental difficulty arises for our approach. In this regard, we analyze the relationship

[2]Technically speaking, the computed values in this experiment are the empirical expectation which can contain a statistical error. However, since we have no access to the data generating distribution, we perform the analysis as if these values are the population expectations.