Adjustment and Alignment for Unbiased Open Set Domain Adaptation (*Supplementary Material*)

Wuyang Li[1] Jie Liu[1] Bo Han[2] Yixuan Yuan[3,*]
[1]City University of Hong Kong [2]Hong Kong Baptist University [3]The Chinese University of Hong Kong
{wuyangli2,jliu.ee}-c@my.cityu.edu.hk bhanml@comp.hkbu.edu.hk yxyuan@ee.cuhk.edu.hk

In this supplementary material, we provide more theoretical and experimental justifications with four sections.

- Sec. A: causal explanation on the bias and its solutions;
- Sec. B: the theoretical proof of front-door adjustment and the deployed OSDA adjustment;
- Sec. C: more discussions and clarifications;
- Sec. D: sensitivity analysis on the proposed method.

A. Justifying the Bias

A.1. Formulating the Bias with Causality

As shown in Figure 1(a), the open-set context C is the confounder [6] that serves as the shared *cause* (head) in two causal links $C \rightarrow X$ and $C \rightarrow Y$ (highlighted in blue), leading to a biased observation $P(Y|X)$. This bias is caused by the confounding effect of C [6], and is well-proven in [6], and is theoretically grounded in various computer vision tasks [5, 8–12].

A.2. Solving the Bias with Causality

To address the bias, we follow the causal theory [6] to deploy the *do*-calculus [6], which corrects the biased posterior $P(Y|X)$ with $P(Y|do(X))$. Pearl and Mackenzie [6] have given two available solutions to implement the *do*-calculus from the theoretical perspective, *i.e.*, backdoor adjustment (Figure 1(b)) and front-door adjustment (Figure 1(c)).

Backdoor Adjustment. Existing works [9, 12] conduct backdoor adjustment [6] (see Figure 1(b)) for debiasing, which aims to cut the link $C \rightarrow X$ to remove the confounding effect [6] of C. Specifically, as for the implementation, they [9, 12] decouple the context C into bias-related components $C = \{C_1, C_2, ...C_n\}$ via dataset-level statistics and use them to guide model training. For example, with ground-truth labels, VC-RCNN [9] calculates class centers by averaging per-class samples as C_n, generating a context dictionary $\in \mathbb{R}^{K \times D}$ (K is the class number and D is the channel dimension.). Then, they use the context dictionary to guide the learning of X through soft weighing, which introduces balanced class knowledge in each training sample. However, in OSDA, the non-available novel-class images and labels in the source domain lead to an **unobservable** open-set context, making it inevitable to deploy the backdoor adjustment [6] based debiasing strategy.

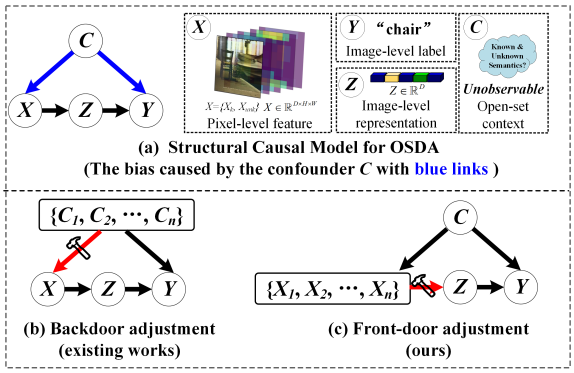

Figure 1. Illustration of (a) the proposed OSDA Structural Causal Model, (b) backdoor adjustment, and (c) front-door adjustment.

Front-Door Adjustment. This work breaks through this barrier by implementing the unexplored front-door adjustment, which allows the unobservable counfounder [6], as shown in Figure 1(c). Instead of cutting the link $C \rightarrow X$ through decoupling the context, the front-door adjustment cuts the link $X \rightarrow Z$ to remove the confounding effect of C [6]. Hence, our crucial insight of deploying front-door adjustment lies in decoupling $X = \{X_1, X_2, ..., X_n\}$, which is implemented as the decoupled base-class and novel-class regions $X = \{X_b, X_n\}$ for unbiased OSDA.

B. Theoretical Proofs

B.1. Proof of the Front-Door Adjustment (Eq. 1 in the Main Paper)

Following [6], the proof of the original front-door adjustment considers two causal rules [6].

*Corresponding author.

This work was supported by Hong Kong Research Grants Council (RGC) General Research Fund 11211221, and Innovation and Technology Commission-Innovation and Technology Fund ITS/100/20.

- **Causal Rule 2**: $P(Y|do(X),Z) = P(Y|X,Z)$ *if* Z *satisfies the back-door criterion*
- **Causal Rule 3**: $P(Y|do(X)) = P(Y)$ *if there is no path from* X *to* Y *with only forward-directed arrows*

Note that Rule 2 and Rule 3 have been proven satisfied in the proposed causal model [6]. Then, the proof consists of the following steps [6],

$$
\begin{aligned}
&P(Y|do(X)) \\
=&\sum_{Z} P(Y|do(X),Z)P(Z|do(X)) \\
&(\textbf{Probability Axioms}) \\
=&\sum_{Z} P(Y|do(X),do(Z))P(Z|do(X)) \\
&(\textbf{Causal Rule 2}) \\
=&\sum_{Z} P(Y|do(X),do(Z))P(Z|X) \\
&(\textbf{Causal Rule 2}) \\
=&\sum_{Z} P(Y|do(Z))P(Z|X) \\
&(\textbf{Causal Rule 3}) \\
=&\sum_{X'\subseteq X}\sum_{Z} P(Y|do(Z),X')P(X'|do(Z))P(Z|X) \\
&(\textbf{Probability Axioms}) \\
=&\sum_{X'\subseteq X}\sum_{Z} P(Y|Z,X')P(X'|do(Z))P(Z|X) \\
&(\textbf{Causal Rule 2}) \\
=&\sum_{X'\subseteq X}\sum_{Z} P(Y|Z,X')P(X')P(Z|X), \\
&(\textbf{Causal Rule 3})
\end{aligned}
\tag{1}
$$

where $X' \subseteq X$ indicates the decoupled components [6], which is formulated as $X = \{X_b, X_n\}$ in this paper.

B.2. Proof of the Deployed OSDA Adjustment (Eq. 3 in the Main Paper)

With the inherent and decoupled base-class and novel-class regions $X = \{X_b, X_n\}$ in an image, we can open the summation symbol in Eq. 1 for unbiased OSDA, which is denoted as follows,

$$
\begin{aligned}
P(Y|do(X)) = &P(Y|Z_b,X_b)P(X_b)P(Z_b|X) \\
&+ P(Y|Z_n,X_n)P(X_n)P(Z_n|X) \\
&+ P(Y|Z_n,X_b)P(X_b)P(Z_n|X) \\
&+ P(Y|Z_b,X_n)P(X_n)P(Z_b|X).
\end{aligned}
\tag{2}
$$

Since the global average pooling $X \rightarrow Z$ doesn't change the semantic role of an image [3] in the deep-learning-based image recognition, we have $P(Z_b|X_b) = P(Z_n|X_n) = 1$ and $P(Z_b|X_n) = P(Z_n|X_b) = 0$, which can be used to justify some key items in Eq. 2. On the one hand, according to Total Probability Theorem, the conditional probability $P(Z_{b/n}|X)$ can be rewritten as follows (we use blue and red color to highlight the zero and one value term),

$$
\begin{aligned}
P(Z_b|X) &= P(Z_b|X_b)P(X_b|X) + P(Z_b|X_n)P(X_n|X) \\
&= P(X_b|X) \\
P(Z_n|X) &= P(Z_n|X_b)P(X_b|X) + P(Z_n|X_n)P(X_n|X) \\
&= P(X_n|X)
\end{aligned}
\tag{3}
$$

One the other hand, the joint probability $P(Z_{b/n}, X_{b/n})$ can be rewritten with Bayes' Rule, denoted as follows,

$$
\begin{aligned}
P(Z_b,X_b) &= P(Z_b|X_b)P(X_b) = P(X_b); \\
P(Z_n,X_n) &= P(Z_n|X_n)P(X_n) = P(X_n); \\
P(Z_b,X_n) &= P(Z_b|X_n)P(X_n) = 0; \\
P(Z_n,X_b) &= P(Z_n|X_b)P(X_b) = 0.
\end{aligned}
\tag{4}
$$

From the above analysis, we have the following explanation in terms of $P(Y|Z_{b/n}, X_{b/n})$. Firstly, optimizing $P(Y|Z_b, X_b)$ and $P(Y|X_b)$ are equivalent for the model learning since the conditioned events are equivalent: $P(Z_b, X_b) = P(X_b)$. This is also satisfied between $P(Y|Z_n, X_n)$ and $P(Y|X_n)$ with $P(Z_n, X_n) = P(X_n)$. Secondly, for the conditional probability $P(Y|Z_b, X_n)$ and $P(Y|Z_n, X_b)$, we can observe that they are conditioned on an event of probability zero: $P(Z_b, X_n) = P(Z_n, X_b) = 0$, which can be approximated with a non-informative constant [2]. Thus, these two items don't contribute to making the class-level decision [2] for image recognition due to its non-discriminative property. Finally, after introducing Eq. 3 and Eq. 4 into Eq. 2, we have the simplified and deployable front-door adjustment to achieve an unbiased OSDA:

$$
\begin{aligned}
P(Y|do(X)) = &P(Y|X_b)P(X_b)P(X_b|X) \\
&+ P(Y|X_n)P(X_n)P(X_n|X).
\end{aligned}
\tag{5}
$$

Hence, the key insight to correct the biased learning lies in considering both base-class and novel-class posterior $P(Y|X_b)$ and $P(Y|X_n)$, which is achieved by the proposed Front-Door Adjustment loss $\mathcal{L}_{FDA}$.

C. Discussion and Clarification

Why ambiguous samples generate offsetting signals in the two heads of DCA. For ambiguous (uncertain) samples, the signals of two heads of the proposed DCA module tend to be **comparable** due to balanced mask entries $\mathbf{M}_b^i \approx \mathbf{M}_n^i$. Note that two heads adapt samples to base and novel distribution **orthogonally and respectively**. Hence, the comparable intensity of the two heads tends to prevent incorrect adaptation to either distribution.

Explanation on FDA (Eq.5) and DCA (Eq.7) in the main paper. 1) The relation between Eq.5/7. Though both equations aim to discover base and novel regions $X_{b/n}$, Eq.5 requires image *labels* (see Eq.4) to discover labeled and unlabeled parts $X_b = \{X_{lb}, X_{ub}\}$. Differently, Eq.7 works in a *label-free* manner to find X_b and X_n in both domains, which cannot separate X_{lb} and X_{ub}. **2) Why does DCA not consider unlabeled regions?** DCA generates base-class masks $\mathbf{M}_b^i$ (Eq.7) to align $P(X_b) = P(X_{lb}, X_{ub})$, which has considered labeled X_{lb} and unlabeled regions X_{ub}. Note that DCA does not need to separate X_{lb} and X_{ub}, since it aligns the whole base-class distribution $P(X_b)$ instead of $P(X_{lb/ub})$.

Strategy to prevent a risky selection in FDA. We first follow Fig.4 to justify $X_{lb/ub/n}$ in images, and then select X_n to generate the loss when there are more **recognized** X_{lb} than X_n. The reason is that very limited X_{lb} can be recognized in the early training stage, hence, most regions will be wrongly assumed as X_n. This strategy avoids wrongly assuming most samples as X_n, yielding a sub-optimal (risky) selection. Figure 4 of the main paper illustrates the basic idea of region selection instead of the whole learning procedure. With the model training, the number of $X_{lb/ub}$ will increase, and X_n will decline due to more recognized $X_{lb/ub}$, which is shown in Table 1.

Epoch	0	1	2	3	4
$\lvert X_{lb}\rvert : \lvert X_n\rvert$	0: 49	9: 32	17:22	17:16	16:14

Table 1. Statistics of the ratio between X_{lb} and X_n

D. Sensitivity Analysis

Sensitivity on the feature resolution. Considering that ANNA relies on fine-grained representation, we explore the effect on the block resolution (Table 2) by changing corresponding input image scales. With an increased feature resolution, more fine-grained visual blocks can be obtained in each image to remove the bias. Our method can improve further with the increase of feature resolution, *e.g.*, giving a better 77.8% HOS (9×9) compared with 76.8% HOS (7×7), showing its great potential in scene understanding.

Resolution	Ar→Rw			Cl→Pr		
	OS*	UNK	**HOS**	OS*	UNK	**HOS**
7×7	74.1	79.7	76.8	64.2	73.6	68.6
8×8	76.9	77.7	77.3	**64.7**	76.0	69.9
9×9	76.4	**79.9**	**77.8**	63.4	**78.6**	**70.1**
10×10	**80.1**	74.1	77.0	64.2	76.9	70.0

Table 2. Analysis on the fine-grained visual block resolutions. A larger resolution indicates using more blocks for unbiased OSDA.

Sensitivity on the multiple runs. As shown in Table 3. we further report the experimental comparison with three runs as [1] on HOS (%) to evaluate the model robustness. It can be observed that the proposed method is able to achieve the most robust performance with 70.6±0.2, compared with the biased counterparts, *e.g.*, OSBP with 64.9 ± 0.5, due to satisfactory semantic perception with the debiasing effects. Moreover, our method also surpasses all the counterparts by a large margin in terms of the average performance among several runs, indicating the promising effects of the proposed theoretically grounded method.

STA$_{sum}$ [4]	STA$_{max}$ [4]	OSBP [7]	ROS [1]	Anna (ours)
62.1± 2.3	61.0 ± 0.5	64.9 ± 0.5	66.5 ± 0.3	70.6 ± 0.2

Table 3. Comparison results with three runs on Office-Home.

References

[1] Silvia Bucci, Mohammad Reza Loghmani, and Tatiana Tommasi. On the effectiveness of image rotation for open set domain adaptation. In *ECCV*, pages 422–438. Springer, 2020. 3

[2] William K Goosens. Alternative axiomatizations of elementary probability theory. *Notre Dame Journal of Formal Logic*, 20(1):227–239, 1979. 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[4] Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *CVPR*, pages 2927–2936, 2019. 3

[5] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*, pages 12700–12710, 2021. 1

[6] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018. 1, 2

[7] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *ECCV*, pages 153–168, 2018. 3

[8] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020. 1

[9] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *CVPR*, pages 10760–10770, 2020. 1

[10] Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *ICLR*, 2022. 1

[11] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *NeurIPS*, 2020. 1

[12] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *NeurIPS*, 33, 2020. 1