# Adjustment and Alignment for Unbiased Open Set Domain Adaptation

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

## Abstract

*Open Set Domain Adaptation (OSDA) transfers the model from a label-rich domain to a label-free one containing novel-class samples. Existing OSDA works overlook abundant novel-class semantics hidden in the source domain, leading to a biased model learning and transfer. Although the causality has been studied to remove the semantic-level bias, the non-available novel-class samples result in the failure of existing causal solutions in OSDA. To break through this barrier, we propose a novel causality-driven solution with the unexplored front-door adjustment theory, and then implement it with a theoretically grounded framework, coined Adjustment and Alignment (ANNA), to achieve an unbiased OSDA. In a nutshell, ANNA consists of Front-Door Adjustment (FDA) to correct the biased learning in the source domain and Decoupled Causal Alignment (DCA) to transfer the model unbiasedly. On the one hand, FDA delves into fine-grained visual blocks to discover novel-class regions hidden in the base-class image. Then, it corrects the biased model optimization by implementing causal debiasing. On the other hand, DCA disentangles the base-class and novel-class regions with orthogonal masks, and then adapts the decoupled distribution for an unbiased model transfer. Extensive experiments show that ANNA achieves state-of-the-art results. The code is available at https://github.com/CityU-AIM-Group/Anna.*

## 1. Introduction

Unsupervised Domain Adaptation (UDA) [5, 8, 11, 13] has been well studied to transfer a model from a labeled domain to an unlabeled novel one, notably saving the labeling labor for model re-implementation. However, existing UDA research follows a strong assumption that the two domains must share the same class space, which cannot make correct predictions for novel-class samples. This severely limits real-world applications [25, 29], *e.g.*, product recommendation and pathology identification with unseen classes.

Aiming at addressing this issue, Open Set Domain Adaptation (OSDA) [3, 17, 20, 29, 35] has been studied, which also needs to recognize the novel-class samples in the target domain as *unknown*. As shown in Figure 1(a) (top), following a similar pipeline, most existing works [3, 17, 20, 29, 35] utilize labeled base-class data to train a closed-set classifier in the source domain. Then, in the target domain, they adjust the model with two objectives, *i.e.*, exploring novel samples to achieve base/novel-class separation (novel-class detection) and adapting the base-class distribution (domain alignment). Based on this pipeline, these works can successfully recognize some novel samples in the unlabelled target domain and align the base-class distribution well.

While achieving great success, existing works [17, 20, 29] only consider base-class semantics in the source domain, ignoring the novel-class spreading everywhere. This leads to a semantic-level bias between the base and novel class, further yielding a biased domain transfer for OSDA. To explore the deficiency of this bias, we visualize the base/novel-class activated regions, as shown in col. 1-2 of Figure 1(a) (bottom). It can be observed that existing approaches can successfully find the base-class regions consistent with the image-level ground-truth *chair*, but cannot discover novel-class semantics, *e.g.*, the *yacht*, *sea*, and *ground*, etc. (The base and novel regions are highlighted in Figure 1(c) for better view.) Further, we conduct a per-pixel prediction on deepest features without global average pooling (col. 3), illustrating that the novel regions are misclassified as some non-correlated base classes. These observations imply that this semantic-level bias severely affects the judgment of the classifier even though the classifier can give a correct prediction for the whole image.

Recently, several causality-based approaches [36, 44, 45] have been proposed to solve the semantic-level bias in the closed-set setting. These works [36, 44, 45] first conduct per-class statistics over the whole dataset to decouple the *context*, and then use decoupled components to correct the biased model training in a class-balanced manner. This causal solution can successfully avoid biased model learning since

*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.

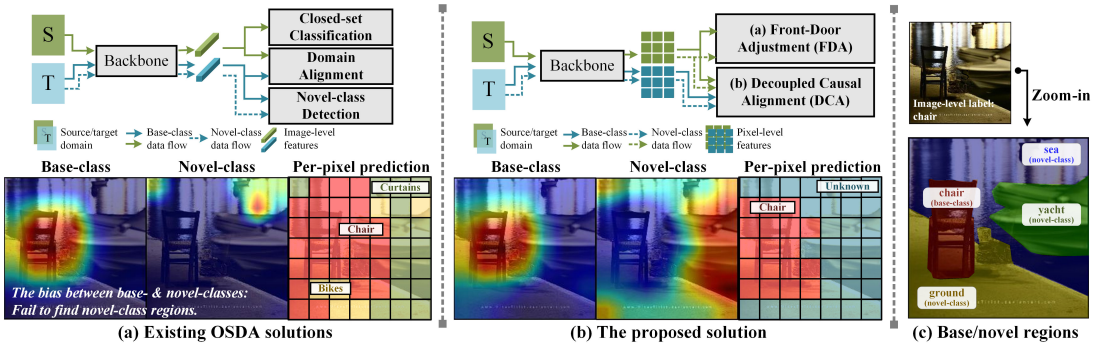


Figure 1. Illustration of the general pipeline (top) and observed bias (bottom) with the base/novel-class activation and per-pixel prediction (we conduct dense classification on each pixel of the 7×7 ResNet-50 [9] feature and highlight the pixels with the same result in the same color.) for (a) existing OSDA approaches, (b) our solution, and (c) base and novel regions in each image.

the knowledge of all classes contributes to training each sample. Hence, the rational idea is to explore the causality to solve the newly observed OSDA bias.

However, it is intractable to implement existing causal solutions [36, 45] in OSDA since the *context* is unobservable in open-set setting [42, 43]. Existing works [36, 45] use backdoor adjustment theory [26] to remove the bias, which relies on the observable context with available data samples. Differently, in OSDA, the context is unobservable [42] since novel-class samples are missing in the source domain [29] and labels are non-available in the target domain, leading to the failure [42] of existing backdoor solutions [36, 44, 45]. Although the front-door adjustment [26] can break through this unobservable dilemma [26] by decoupling data samples instead of context[1], it is still tricky to implement a semantic-level decoupling [36, 45] on each data sample since each image is only assigned a single class label in classification [9]. Fortunately, as shown in Figure 1(c), we observe that each image can be decoupled into base/novel-class regions in this open-set setting, which motivates us to use the **unexplored yet effective** front-door adjustment [26] to remove the bias. Thus, we aim to correct the biased learning in the source domain and then align the decoupled cross-domain distribution to achieve unbiased OSDA. See Sec. 3 for a theoretical analysis with Structural Causal Model.

To address the problems mentioned above, we propose a theoretically grounded framework, Adjustment and Alignment (ANNA) for OSDA (see Figure 1(b) (top)) with causality, which consists of Front-Door Adjustment (FDA) to address the biased learning in the source domain, and Decoupled Causal Alignment (DCA) to transfer the model to the target domain unbiasedly. Specifically, in each base-class image, FDA delves into fine-grained visual blocks to discover novel-class regions, serving for correcting biased model learning with causal adjustment. As for the DCA module, we disentangle cross-domain images into base-class and novel-class regions with orthogonal masks, and then align the decoupled distribution free of bias. As shown in Figure 1(b) (bottom), after eliminating the OSDA bias, the model can capture labeled base-class regions (col. 1) and unlabeled novel-class regions (col. 2) well. Besides, the per-pixel prediction (col. 3) gives a closer look at model inference, showing that ANNA fully considers fine-grained novel semantics like humans before making an image-level prediction. Our main contributions are as follows,

- This work represents the first attempt that observes and formulates the ever-overlooked semantic-level bias in OSDA. To address this issue, we propose a theoretically grounded framework, Adjustment and Alignment (ANNA) with causality, achieving an unbiased OSDA.
- We propose a Front-Door Adjustment (FDA) module to correct the biased closed-set learning, discovering and fully using novel-class regions hidden in images.
- We design a Decoupled Causal Alignment (DCA) to achieve an unbiased model transfer, which decouples cross-domain images with fine-grained regions and aligns the decoupled distribution unbiasedly.
- Extensive experiments on three benchmarks verify that ANNA achieves state-of-the-art performance. ANNA achieves the best HOS on all 12 sub-tasks of the challenging Office-Home benchmark.

## 2. Related Work

**Closed-Set Domain Adaptation (CSDA).** CSDA transfers the model from a labeled source domain to an unlabeled counterpart with a shared class space. Some works [4, 5, 8, 14, 19, 38, 46] deploy tailor-designed discriminators to align the distribution via adversarial learning [8]. Some works [11–13, 15, 31] measure domain discrepancy and adapt via metric learning. Moreover, various self-training techniques [18, 21] are developed to discover reliable target-domain samples for better semantic discriminability. Due the shared class space, CSDA models cannot recognize novel classes with limited real-world applications.

[1] See supplementary materials for a more detailed explanation.

**Open Set Domain Adaptation (OSDA).** OSDA is the extension of CSDA that allows the target domain to contain novel classes. Saito *et al*. [29] raises the OSDA setting that novel-class samples only appear in the target domain for practical application. Recent advances mainly focus on two streams of the research, *i.e*., 1) separating the known and unknown samples in the unlabeled target domain with score-based similarity [17,20], adversarial learning [10,29], rotation-invariance [3], binary classification [17], and 2) aligning the cross-domain known distribution with adversarial alignment [17], pseudo-labeling [20, 35]. However, existing approaches ignore the biased learning in the source domain with limited novel-class discriminability, which is addressed by the proposed causality-driven solutions.

**Causality-based Debiasing.** Causality has been widely studied in various computer vision tasks [23, 24, 32, 32, 36, 37, 39, 42–44] to correct the biased model learning. Most existing works [36, 39, 44, 45] rely on the backdoor adjustment theory [26] for debiasing, which decouples the confounder [26], *e.g*., context [36, 45] at the dataset level and use decoupled components to correct the biased learning [36,45]. Wang *et al*. [36] explore the bias caused by the object co-occurrence and then solve it via backdoor adjustment [26]. Zhang *et al*. [45] implement the backdoor adjustment [26] in weakly-supervised semantic segmentation to generate unbiased pseudo-masks. These works [36, 45] address the bias in closed-set with deployable backdoor adjustment [26], relying on the observable context with available samples. Differently, due to the unobservable context in OSDA, we leverage the unexplored front-door adjustment [26] theory (see Figure 2) to address the biased OSDA.

## 3. Preliminaries and Motivation

**Problem Formulation.** In OSDA, we have labeled source data $\{I_s^i, Y_s^i\}_{i=1}^{N_s}$ and unlabeled target data $\{I_t^i\}_{i=1}^{N_t}$ drawn from inconsistent distribution $P_s \neq P_t$. Both domains share $K$ base classes, while the target domain also contains extra $K'$ target-private novel classes, which are uniformly considered as *unknown* [29] (class $K+1$). OSDA aims to recognize both base/novel classes in the target domain by solving the *1) Open-set* and *2) Cross-domain* challenge [3].

**Structural Causal Model.** In OSDA, we formulate the causality among the pixel-level feature $X$, image-level representation $Z$, image-level label $Y$ and unobservable open-set context $C$ [36] in Figure 2. The node indicates the causal variable, and the directed edge is a specific operation with the causality from the cause (head) to the effect (tail).

$X \rightarrow Z \rightarrow Y$ indicates the causality of image recognition. Given $X$ extracted from the feature extractor, the deep model first abstracts image-level representation $Z$ through pooling [9] ($X \rightarrow Z$), and then separates the high-level embedding into different classes with a classifier ($Z \rightarrow Y$).

$X \leftarrow C \rightarrow Y$. The unobservable context prior $C$ determines the composition of each image [36, 45] ($X \leftarrow C$), *e.g*., putting a *chair* (base-class) next to a *yacht* (novel-class). Moreover, $C$ determines the label distribution with a specific class prior [36, 45], justifying $C \rightarrow Y$.

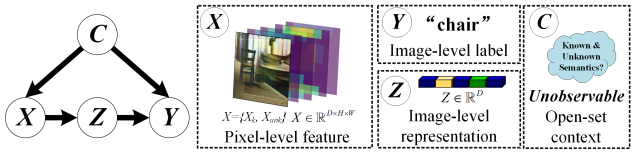


Figure 2. The proposed Structural Causal Model for OSDA.

Hence, as the observation in [36, 45], open-set context $C$ prevents the unbiased learning from $X$ to $Y$ [26]. To break through this barrier, we implement the front-door adjustment [26] with *do*-calculus [26] as follows, (see supplementary materials for the proof.)

$$P(Y|do(X)) = \sum_{X' \subseteq X} \sum_{Z} P(Y|Z, X')P(X')P(Z|X). \tag{1}$$

Then, we ground the adjustment formula in OSDA, which considers the base and novel regions $X = \{X_b, X_n\}$ in each image (see Figure 1(c)). With the link of pooling operation ($X \rightarrow Z$), we also have $Z = \{Z_b, Z_n\}$. Thus, the summation symbols in Eq. 1 can be opened:

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

Based on the proposed causal formulation, we can naturally have $P(Z_b|X_b) = P(Z_n|X_n) = 1$ and $P(Z_b|X_n) = P(Z_n|X_b) = 0$, due to the fact that the pooling operation ($X \rightarrow Z$) won't change the semantic-level role of $X$ [16]. Then, we have, (see supplementary materials for the proof.)

$$\begin{aligned} P(Y|do(X)) = & P(Y|X_b)P(X_b)P(X_b|X) \\ & + P(Y|X_n)P(X_n)P(X_n|X), \end{aligned} \tag{3}$$

where $P(X_{b/n})$ is the marginal distribution of base/novel-class at the dataset level, $P(X_{b/n}|X)$ measures the ratio of base/novel-class semantics for given images, $P(Y|X_{b/n})$ is the optimized posterior for an unbiased classifier.

**Remark 1. Open-set challenge.** With Eq. 3, we can observe that existing works [3, 17, 29] wrongly assume $X = X_b$ and optimize $P(Y|X_b)$ with a closed-set classification (Figure 1(a)) in the source domain. Thus, to address this bias, we propose FDA (Sec. 4.1) to optimize $P(Y|do(X))$. FDA first discovers novel regions $X_n$ in each image $X$ to make $P(Y|X_n)$ and $P(X_n|X)$ measurable, and then introduces an unbiased learning objective $\mathcal{L}_{FDA}$ by optimizing $P(Y|do(X))$ with $P(Y|X_b)$ and $P(Y|X_n)$.

**Remark 2. Cross-domain challenge.** Eq. 3 can be further analyzed by introducing the well-proved Covariate Shift [2,

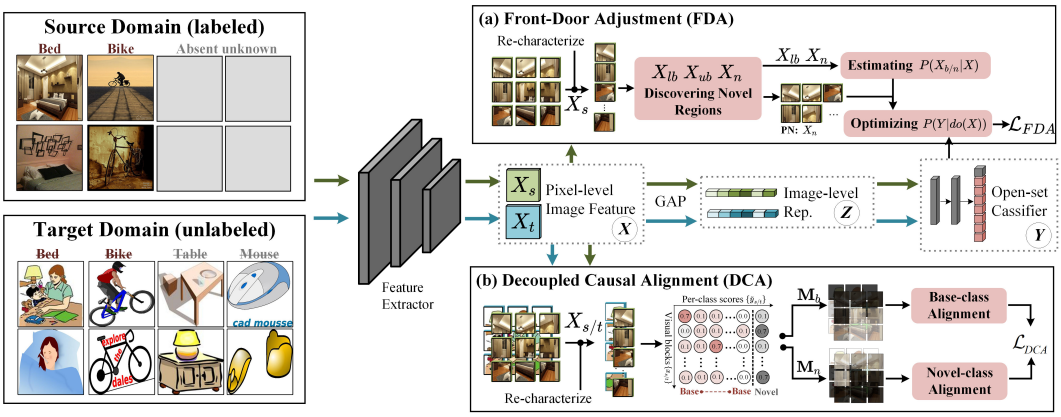


Figure 3. Illustration of the proposed ANNA framework. We utilize $1^{st}$ image in each batch to clarify the workflow of each module. GAP indicates the global average pooling for feature abstraction. Causal variables are highlighted with circles.

30], *i.e.*, different domains have inconsistent marginal distribution $P(X)$ and consistent $P(Y|X)$. Thus, $P(X_{b/n})$, the domain gap will attack, should be aligned to ensure the applicability of unbiased $P(Y|do(X))$ under cross-domain scenarios[2]. Existing OSDA solutions [3, 17, 29] only align base-class distribution $P(X_b)$, failing to achieve an unbiased transfer with misaligned $P(X_n)$. Differently, we propose DCA (Sec. 4.2) to adapt both $P(X_{b/n})$, which generates orthogonal masks to decouple the batch-level observation drawn from $P(X_{b/n})$ for a decoupled adaptation,

## 4. Adjustment and Alignment (ANNA)

The overall workflow is shown in Figure 3. With batch-wise source data $\{I_s^i, Y_s^i\}_{i=1}^B$ and target data $\{I_t^i\}_{i=1}^B$, we use shared feature extractor to obtain pixel-level features $X_{s/t}$, which are sent to (a) Front-Door Adjustment (FDA) and (b) Decoupled Causal Alignment (DCA) for unbiased OSDA. In FDA, we first re-characterize the image feature $X_s$ with fine-grained visual blocks $\{x_s^i\}_{i=1}^N$ and then discover inherent novel regions $X_n = \{x_n^i\}_{i=1}^{N_m}$, serving for implementing $P(Y|do(X))$ with an unbiased learning loss $\mathcal{L}_{FDA}$. DCA transforms cross-domain images $X_{s/t}$ into block-based representation $\{x_{s/t}^i\}_{i=1}^N$, and decouples it into base and novel regions with orthogonal masks $\mathbf{M}_{b/n}$. Then, it aligns the decoupled distribution $P(X_{b/n})$ to remedy the domain gap with an unbiased transfer loss $\mathcal{L}_{DCA}$.

### 4.1. Front-Door Adjustment

With batch-wise source data $\{I_s^i, Y_s^i\}_{i=1}^B$, we extract pixel-level image features $X_s \in \mathbb{R}^{B\times D\times H\times W}$ and send them to FDA. FDA corrects the biased closed-set learning by implementing the front-door formula as Eq. 3.

**Discovering Novel Regions $X_n$.** As modeling $P(Y|X_n)$ in Eq. 3 relies on available novel samples $X_n$, we delve into fine-grained visual blocks to discover intrinsic novel-class regions hidden in base-class images. Specifically, we use visual blocks to re-characterize the image into fine-grained representation $X_s = \{x_s^i\}_{i=1}^N, N = B\times H\times W$, and deploy it on pixel-level features instead of original images to avoid multi-times forward propagation [27, 33]. Then, we categorize the visual blocks into three types to comprehend the image compensation and clarify each type in Figure 4(a).

**(1) Labeled Base regions $X_{lb}$.** We define the main body consistent with image-level labels as LB. According to the labeling habit, data collectors tend to label the commonly seen object with obvious characters, *e.g.*, large scale.

**(2) Unlabeled Base regions $X_{ub}$.** We consider the object in a base class that hasn't been correctly labeled. For example, although *curtain* is a base-class in Office-Home [34], the curtain in $1^{st}$ image hasn't been labeled due to the more notable bed (LB). These blocks harm the feature learning due to its inconsistent semantics with labels.

**(3) Potential Novel regions $X_n$.** Except for LB and UB, all the other blocks could be considered as the PN, *e.g.*, the *door* and *window* in $1^{st}$ image.

Based on our OSDA analysis (Eq. 3), we aim to discover $X_n$ to conduct front-door adjustment. To avoid the influence of $X_{lb}$, the correctly classified highly-confident blocks should be eliminated, since the image classification has induction capacity on label-matched regions. As for removing $X_{ub}$, we utilize the intrinsic cues in the score-ranking, considering the semantic affinity, *i.e.*, the co-occurrent objects[3] have similar semantics compared with non-correlated counterparts [40] in a specific scene. Taking $2^{nd}$ image of Figure 4 as an example, the selected visual block containing *desk-lamp* ($\in X_{ub}$) tends to generate a higher score in the *computer* channel compared with the non-correlated class *bike* in this indoor scene [40].

To this end, we freeze the classifier and forward-propagate all fine-grained visual blocks $\{x_s^i\}_{i=1}^N$ independently to obtain per-class confidence $\{\widetilde{y}_s^i\}_{i=1}^N, \widetilde{y}_s^i \in \mathbb{R}^{K+1}$,

[2]Intuitively, in OSDA, target-private novel-class images [29] tend to result in a larger covariate shift on $P(X_n)$

[3]Each $x \in X_{ub}$ *must* appear with the labeled block in $X_{lb}$ together.

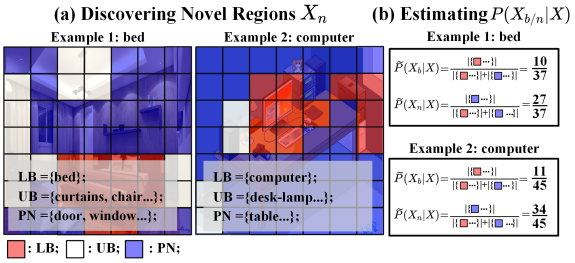


Figure 4. (a) Illustration of the three types of visual blocks, *i.e.*, Labeled Base (LB: $X_{lb}$), Unlabeled Base (UB: $X_{ub}$), and Potential Novel (PN: $X_n$). (b) Toy examples for estimating $P(X_{b/n}|X)$

which are used to discover novel regions $X_n$ as follows,

$$\begin{aligned} X_{lb} &= \{x_s | \text{argmax}(\widetilde{y}_s) = \hat{y}_s\}, \\ X_{ub} &= \{x_s | \hat{y}_s \in \text{Top-K}(\widetilde{y}_s) \wedge \text{argmax}(\widetilde{y}_s) \neq \hat{y}_s\}, \\ X_n &= \{x_s | x_s \notin X_{lb} \wedge x_s \notin X_{ub}\}, \end{aligned} \tag{4}$$

where $\hat{y}_s$ is the image-level label for the selected local visual block. Considering the poor semantic in the early training stage, we only utilize the discovered blocks $X_n$ satisfying $|X_n| < |X_{lb}|$ to prevent the risky selection.

**Estimating $P(X_{b/n}|X)$.** To measure the $P(X_{b/n}|X)$ in Eq. 3, we estimate the probability $\widetilde{P}(X_{b/n}|X)$ in each training batch by counting the number of LB and PN blocks as follows, (see Figure 4(b) for two toy examples.)

$$\begin{aligned} \eta_b &= \widetilde{P}(X_b|X) = \frac{|X_{lb}|}{|X_{lb}| + |X_n|}, \\ \eta_n &= \widetilde{P}(X_n|X) = \frac{|X_n|}{|X_{lb}| + |X_n|}, \end{aligned} \tag{5}$$

where $\widetilde{P}(X_{b/n}|X)$ is the estimated conditional probability for each training batch. It is worth noting that we ignore $X_{ub}$ to avoid the influence of inconsistent semantics.

**Unbiased Learning by Optimizing $P(Y|do(X))$.** Then, based on the discovered novel visual blocks $X_n$, we can naturally implement the unbiased optimization objective (consistent with Eq. 3) as follows,

$$\mathcal{L}_{FDA} = \eta_b \mathcal{L}_b + \eta_n \mathcal{L}_n, \tag{6}$$

where $\eta_{b/n} = \widetilde{P}(X_{b/n}|X)$, $\mathcal{L}_b$ is implemented with the standard closed-set classification loss [29] for $P(Y|X_b)$ and $\mathcal{L}_n = -\frac{1}{|X_n|}\sum_{x_n \in X_n} log(p(\widetilde{y}_n = K+1|x_n))$ aims to maximize the novel-class probability on $X_n$ for $P(Y|X_n)$. Moreover, for the terms $P(X_{b/n})$ in Eq. 3, we follow [36, 45] to use an evenly distributed approximation with balanced constant entries $P(X_{b/n}) = 0.5$, thereby, independent of the model optimization. Considering that the feature extractor may overfit some non-informative backgrounds, we add a gradient scaling layer after the feature extractor as [29] for $\mathcal{L}_n$, which scales the gradient with a constant $\lambda \in [0, 1]$ during back-propagating $\mathcal{L}_n$ to relieve such influence. (See Sec.5.4 for experimental analysis.)

## 4.2. Decoupled Causal Alignment

After correcting the biased learning in the source domain, DCA is proposed to transfer the model to the target domain unbiasedly. DCA first decouples each image $X$ into base and novel class regions through generating orthogonal masks $\mathbf{M}_{b/n}$ and then aligns the decoupled distribution to achieve an unbiased cross-domain transfer.

**Orthogonal Mask Generation.** To estimate and align the decoupled distribution unbiasedly, we propose a set of orthogonal masks to decouple base and novel class regions in each image. Specifically, we first transform source and target image features $X_{s/t} \in \mathbb{R}^{B\times D\times H\times W}$ into fine-grained blocks $X_{s/t} = \{x^i_{s/t}\}^N_{i=1}, N = B \times H \times W$. Considering the abundant discriminative cues hidden in the output space [5, 38], we freeze the classifier and forward propagate $\{x^i_{s/t}\}^N_{i=1}$ independently to obtain the per-class predictions $\{\widetilde{y}^i_{s/t}\}^N_{i=1}$ for $K+1$ classes. Then, we establish the base-class mask $\mathbf{M}_b$ by looking at the first $K$ channels (for base classes) and generate the novel-class mask $\mathbf{M}_n$ with the $K+1$-th channel(for *unknown*) as follows,

$$\begin{aligned} \mathbf{M}^i_{b,s/t} &= \text{Detach}(\sum_{k=1}^{K} p(\widetilde{y}^i_{s/t} = k|x^i_{s/t})), \\ \mathbf{M}^i_{n,s/t} &= \text{Detach}(p(\widetilde{y}^i_{s/t} = K+1|x^i_{s/t})), \end{aligned} \tag{7}$$

where $\text{Detach}(\cdot)$ is a gradient detach operation to prevent the gradient back-propagation, and $p(\widetilde{y}^i_{s/t} = k|x^i_{s/t})$ indicates the confidence of class $k$ for the visual block $x^i_{s/t}$. Hence, the block with higher base-class confidence tends to generate a larger entry in $\mathbf{M}_b$ while giving a smaller value for $\mathbf{M}_n$, and vice versa. This simple yet effective mechanism can decouple the fine-grained regions in both domains.

**Unbiased Transfer by Aligning $P(X_{b/n})$.** Based on the decoupled regions, we transfer the model by aligning the cross-domain decoupled distribution. It is worth noting that our method is philosophically different from existing biased OSDA approaches [3, 17, 20, 25, 35], since they only align base-class distribution by removing the novel-class samples. Specifically, we implement a double-head discriminator to align decoupled fine-grained regions, which consists of a base-class head $f_b(\cdot)$ and a novel-class head $f_n(\cdot)$ followed by a binary cross-entropy loss:

$$\begin{aligned} \mathcal{L}_{DCA} = &-\sum_{i=1}^{|X_s|}\sum_{o}^{\{b,n\}} \mathbf{M}^i_{o,s} D log(f_o(x^i_s)) \\ &-\sum_{i=1}^{|X_t|}\sum_{o}^{\{b,n\}} \mathbf{M}^i_{o,t}(1-D) log(f_o(x^i_t)), \end{aligned} \tag{8}$$

where $D$ is the domain label, $\mathbf{M}_{b/n}$ is the orthogonal mask. With the guidance of $\mathbf{M}_{b/n}$, two decoupled components focus on independent regions to adapt the $P(X_{b/n})$ respec-

| Office-Home | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | E2E | Ar→Cl | | | Ar→Pr | | | Ar→Rw | | | Cl→Ar | | | Cl→Pr | | | Cl→Rw | | |
| | | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** |
| $STA_{sum}$ [17] | ✗ | 50.8 | 63.4 | 56.3 | 68.7 | 59.7 | 63.7 | 81.1 | 50.5 | 62.1 | 53.0 | 63.9 | 57.9 | 61.4 | 63.5 | 62.5 | 69.8 | 63.2 | 66.3 |
| $STA_{max}$ | ✗ | 46.0 | 72.3 | 55.8 | 68.0 | 48.4 | 54.0 | 78.6 | 60.4 | 68.3 | 51.4 | 65.0 | 57.4 | 61.8 | 59.1 | 60.4 | 67.0 | 66.7 | 66.8 |
| OSBP [29] | ✓ | 50.2 | 61.1 | 55.1 | 71.8 | 59.8 | 65.2 | 79.3 | 67.5 | 72.9 | 59.4 | 70.3 | 64.3 | 67.0 | 62.7 | 64.7 | 72.0 | 69.2 | 70.6 |
| UAN [41] | ✗ | 62.4 | 0.0 | 0.0 | **81.1** | 0.0 | 0.0 | **88.2** | 0.1 | 0.2 | 70.5 | 0.0 | 0.0 | **74.0** | 0.1 | 0.2 | **80.6** | 0.1 | 0.2 |
| DAOD [7] | ✓ | **72.6** | 51.8 | 60.5 | 55.3 | 57.9 | 56.6 | 78.2 | 62.6 | 69.5 | 59.1 | 61.7 | 60.4 | 70.8 | 52.6 | 60.4 | 77.8 | 57.0 | 65.8 |
| PGL [20] | ✗ | 63.3 | 19.1 | 29.3 | 78.9 | 32.1 | 45.6 | 87.7 | 40.9 | 55.8 | **85.9** | 5.3 | 10.0 | 73.9 | 24.5 | 36.8 | 70.2 | 33.8 | 45.6 |
| OSLPP [35] | ✗ | 55.9 | 67.1 | 61.0 | 72.5 | 73.1 | 72.8 | 80.1 | 69.4 | 74.3 | 49.6 | **79.0** | 60.9 | 61.6 | 73.3 | 66.9 | 67.2 | 73.9 | 70.4 |
| ROS [3] | ✗ | 50.6 | 74.1 | 60.1 | 68.4 | 70.3 | 69.3 | 75.8 | 77.2 | 76.5 | 53.6 | 65.5 | 58.9 | 59.8 | 71.6 | 65.2 | 65.3 | 72.2 | 68.6 |
| **Ours** | ✓ | 61.4 | **78.7** | **69.0** | 68.3 | **79.9** | **73.7** | 74.1 | **79.7** | **76.8** | 58.0 | 73.1 | **64.7** | 64.2 | **73.6** | **68.6** | 66.9 | **80.2** | **73.0** |

| Method | Pr→Ar | | | Pr→Cl | | | Pr→Rw | | | Rw→Ar | | | Rw→Cl | | | Rw→Pr | | | **Average** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** |
| $STA_{sum}$ | 55.4 | 73.7 | 63.1 | 44.7 | 71.5 | 55.0 | 78.1 | 63.3 | 69.7 | 67.9 | 62.3 | 65.0 | 51.4 | 57.9 | 54.2 | 77.9 | 58.0 | 66.4 | 63.4 | 62.6 | 61.9 |
| $STA_{max}$ | 54.2 | 72.4 | 61.9 | 44.2 | 67.1 | 53.2 | 76.2 | 64.3 | 69.5 | 67.5 | 66.7 | 67.1 | 49.9 | 61.1 | 54.5 | 77.1 | 55.4 | 64.5 | 61.8 | 63.3 | 61.1 |
| OSBP | 59.1 | 68.1 | 63.2 | 44.5 | 66.3 | 53.2 | 76.2 | 71.7 | 73.9 | 66.1 | 67.3 | 66.7 | 48.0 | 63.0 | 54.5 | 76.3 | 68.6 | 72.3 | 64.1 | 66.3 | 64.7 |
| UAN | **73.7** | 0.0 | 0.0 | 59.1 | 0.0 | 0.0 | 84.0 | 0.1 | 0.2 | 77.5 | 0.1 | 0.2 | 66.2 | 0.0 | 0.0 | **85.0** | 0.1 | 0.1 | 75.2 | 0.0 | 0.1 |
| DAOD | 71.3 | 50.5 | 59.1 | 58.4 | 42.8 | 49.4 | 81.8 | 50.6 | 62.5 | 66.7 | 43.3 | 52.5 | 60.0 | 36.6 | 45.5 | 84.1 | 34.7 | 49.1 | 69.6 | 50.2 | 57.6 |
| PGL | **73.7** | 34.7 | 47.2 | **59.2** | 38.4 | 46.6 | **84.8** | 27.6 | 41.6 | **81.5** | 6.1 | 11.4 | **68.8** | 0.0 | 0.0 | 84.8 | 38.0 | 52.5 | **76.1** | 25.0 | 35.2 |
| OSLPP | 54.6 | **76.2** | 63.6 | 53.1 | 67.1 | 59.3 | 77.0 | 71.2 | 74.0 | 60.8 | 75.0 | 67.2 | 54.4 | 64.3 | 59.0 | 78.4 | 70.8 | 74.4 | 63.8 | 71.7 | 67.0 |
| ROS | 57.3 | 64.3 | 60.6 | 46.5 | 71.2 | 56.3 | 70.8 | 78.4 | 74.4 | 67.0 | 70.8 | 68.8 | 51.5 | 73.0 | 60.4 | 72.0 | 80.0 | 75.7 | 61.6 | 72.4 | 66.2 |
| **Ours** | 63.0 | 70.3 | **66.5** | 54.6 | **74.8** | **63.1** | 74.3 | **78.9** | **76.6** | 66.1 | **77.3** | **71.3** | 59.7 | **73.1** | **65.7** | 76.4 | **81.0** | **78.7** | 65.6 | **76.7** | **70.7** |

| Image-CLEF | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | E2E | B→C | | | B→I | | | B→P | | | C→B | | | C→I | | | C→P | | |
| | | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** |
| OSNN [22] | ✓ | 92.3 | 63.0 | 74.9 | 79.6 | 61.0 | 69.1 | 68.3 | 59.3 | 63.5 | 65.3 | 47.3 | 54.9 | 84.3 | 47.0 | 60.4 | 75.3 | 46.3 | 57.3 |
| STA [17] | ✗ | 93.3 | 51.7 | 66.5 | **86.0** | 60.7 | 71.2 | **77.7** | 48.7 | 59.8 | 61.3 | 69.7 | 65.2 | **91.7** | 66.7 | 77.2 | 84.0 | 54.0 | 65.7 |
| OSBP [29] | ✓ | 87.0 | 81.0 | 83.9 | 85.3 | 65.7 | 74.3 | 66.3 | 66.7 | 66.5 | **62.0** | 58.0 | 59.9 | 89.0 | 80.0 | 84.3 | **87.7** | 53.7 | 66.7 |
| DAOD [7] | ✓ | 79.4 | 82.0 | 80.7 | 78.4 | **90.9** | **84.3** | 72.1 | **80.8** | **76.3** | 51.3 | 47.1 | 49.1 | 79.0 | 88.6 | 83.6 | 74.5 | 78.9 | 76.7 |
| ROS [3] | ✗ | 78.3 | 90.0 | 83.8 | 73.0 | 76.3 | 74.6 | 59.0 | 67.3 | 62.9 | 59.0 | 68.3 | 63.3 | 78.3 | 83.0 | 80.6 | 68.7 | 78.7 | 73.3 |
| **Ours** | ✓ | **95.3** | **98.3** | **96.8** | 81.3 | 84.7 | 83.0 | 74.0 | 75.0 | 74.5 | 58.0 | **83.0** | **68.3** | 87.0 | **93.0** | **89.9** | 78.7 | **84.0** | **81.2** |

| Method | I→B | | | I→C | | | I→P | | | P→B | | | P→C | | | P→I | | | **Average** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** |
| OSNN | 62.0 | 41.6 | 49.8 | 92.0 | 41.3 | 57.0 | 81.3 | 40.6 | 54.2 | 55.0 | 49.6 | 52.2 | 86.0 | 55.3 | 67.3 | 82.0 | 52.6 | 64.1 | 77.0 | 50.4 | 60.4 |
| STA | **62.3** | 54.0 | 57.9 | 94.0 | 53.7 | 68.4 | 80.7 | 59.0 | 68.2 | **61.3** | 43.7 | 51.0 | 93.7 | 47.7 | 63.2 | **90.0** | 51.0 | 65.1 | **81.3** | 55.1 | 65.0 |
| OSBP | 55.7 | 60.7 | 58.1 | 80.7 | 92.7 | 86.3 | 66.3 | 74.3 | 70.1 | 52.3 | 61.0 | 56.3 | **94.0** | 68.0 | 78.9 | 66.0 | 80.7 | 72.6 | 74.4 | 70.2 | 71.5 |
| DAOD | 54.5 | 56.9 | 55.7 | 80.3 | 82.0 | 81.2 | 73.3 | 80.8 | 76.9 | 51.7 | 51.0 | 51.3 | 79.0 | 82.0 | 80.5 | 79.6 | **86.6** | 83.9 | 71.1 | 75.8 | 73.3 |
| ROS | 58.0 | 59.7 | 58.8 | 88.7 | 92.7 | 90.6 | 78.0 | 76.0 | 77.0 | 47.3 | 59.3 | 52.7 | 71.3 | 90.3 | 79.7 | 79.7 | 81.3 | 80.5 | 69.9 | 76.9 | 73.1 |
| **Ours** | 56.0 | **78.0** | **65.2** | **94.3** | **97.7** | **96.0** | **80.7** | **82.7** | **81.7** | 54.0 | **73.7** | **62.3** | **94.0** | **93.7** | **93.8** | 85.0 | 83.3 | **84.2** | 78.2 | **85.6** | **81.4** |

Table 1. Comparison results (%) of Office-Home (top) and Image-CLEF (bottom). E2E indicates end-to-end (single-stage) training.

tively. Moreover, the ambiguous samples with balanced values in $\mathbf{M}_b$ and $\mathbf{M}_n$ could encourage to generate offsetting signals in the two heads, relieving the sub-optimal transfer between unreliable base-class and novel-class semantics.

### 4.3. Model Optimization

During the training stage of the proposed ANNA, we implement the unbiased optimization objective as follows,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{FDA} + \lambda_2 \mathcal{L}_{DCA} + \mathcal{L}_{base}, \tag{9}$$

where $\mathcal{L}_{FDA}$ is for the unbiased learning in the source domain, $\mathcal{L}_{DCA}$ is used for the unbiased transfer to the target domain. $\mathcal{L}_{base} = -0.5log(p(\tilde{Y}_t = K+1|Z_t)) - 0.5log(1 - p(\tilde{Y}_t = K+1|Z_t))$ ($Z_t$ is the target-domain image) is a biased baseline [29], based on which we justify the debiasing effect. $\lambda_{1/2}$ are empirically set 1.0 (see Table 4).

## 5. Experiments

### 5.1. Experimental Setup

**Dataset Settings.** Extensive experiments are conducted on three benchmarks following the standard setting [3, 29, 35]. *1) Office-Home* [34] consists of 65 kinds of labeled images deriving from four specific domains, Art (Ar), Clipart (Cl), Product (Pr), and Real World (Rw). We use the first 25 categories in alphabetic order as the known class and the remaining 40 classes as unknown. *2) Image-CLEF* [1] is the sub-set of four large-scale and publicly available databases, including Bing (B), Caltech-256 (C), ImageNet (I) and PASCAL VOC-2012 (P), with 12 shared common classes. We utilize the first 6 classes as the known class and the rest as the unknown with 12 source-target combinations. *3) Office-31* [28] covers three domains, Amazon (A), Dslr (D), and Webcam (W), with 31 classes, where the first 10 classes are known and the last 11 classes are unknown.

**Evaluation Metrics.** Following the main OSDA stream [3, 17, 29, 35], we use three evaluation metrics to compare the performance, including the average class accuracy over known classes (OS*), the accuracy of unknown class (UNK), and the harmonic mean of OS* and UNK (HOS $= 2\frac{\text{OS}^* \times \text{UNK}}{\text{OS}^* + \text{UNK}}$ [3]). HOS can measure the unbiased property and has been considered as the core metric in the latest OSDA literature [3, 35] since it requires working well on both base and novel classes in an unbiased manner.

**Implementation Details.** All experiments are conducted with the ImageNet [6] pre-trained ResNet-50 [9] feature extractor with an 224×224 input scale as [29]. Each head in DCA consists of a Gradient Reversal Layer [8], two stacked

| Method | E2E | Office-31 | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A→D | | | A→W | | | D→A | | | D→W | | | W→A | | | W→D | | | Average | | |
| | | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** |
| $STA_{sum}$ [17] | ✗ | 95.4 | 45.5 | 61.6 | 92.1 | 58.0 | 71.0 | **94.1** | 55.0 | 69.4 | 97.1 | 49.7 | 65.5 | 92.1 | 46.2 | 60.9 | 96.6 | 48.5 | 64.4 | **94.6** | 50.5 | 65.5 |
| $STA_{max}$ [17] | ✗ | 91.0 | 63.9 | 75.0 | 86.7 | 67.6 | 75.9 | 83.1 | 65.9 | 73.2 | 94.1 | 55.5 | 69.8 | 66.2 | 68.0 | 66.1 | 84.9 | 67.8 | 75.2 | 84.3 | 64.8 | 72.6 |
| OSBP [29] | ✓ | 90.5 | 75.5 | 82.4 | 86.8 | 79.2 | 82.7 | 76.1 | 72.3 | 75.1 | 97.7 | 96.7 | 97.2 | 73.0 | 74.4 | 73.7 | 99.1 | 84.2 | 91.1 | 87.2 | 80.4 | 83.7 |
| UAN [41] | ✗ | **95.6** | 24.4 | 38.9 | **95.5** | 31.0 | 46.8 | 93.5 | 53.4 | 68.0 | **99.8** | 52.5 | 68.8 | **94.1** | 38.8 | 54.9 | 81.5 | 41.4 | 53.0 | 93.4 | 40.3 | 55.1 |
| OSLPP [35] | ✗ | 92.6 | **90.4** | **91.5** | 89.5 | **88.4** | **89.0** | 82.1 | 76.6 | 79.3 | 96.9 | 88.0 | 92.3 | 78.9 | 78.5 | 78.7 | 95.8 | 91.5 | 93.6 | 89.3 | 85.6 | 87.4 |
| ROS [3] | ✗ | 87.5 | 77.8 | 82.4 | 88.4 | 76.7 | 82.1 | 74.8 | 81.2 | 77.9 | 99.3 | 93.0 | 96.0 | 69.7 | 86.6 | 77.2 | **100.0** | **99.4** | **99.7** | 86.6 | 85.8 | 85.9 |
| **Ours** | ✓ | 93.2 | 76.1 | 83.8 | 82.8 | **88.4** | 85.5 | 75.4 | **91.1** | **82.5** | 99.4 | **99.6** | **99.5** | 76.0 | **87.9** | **81.6** | **100.0** | 96.8 | 98.4 | 87.8 | **90.0** | **88.6** |

Table 2. Comparison results (%) on the Office31 benchmark. E2E indicates end-to-end training.

| Row | FDA | DCA | | Ar→Pr | | | Pr→Cl | | | Cl→Rw | | | Rw→Ar | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | base | novel | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** | OS* | UNK | **HOS** |
| 1 | ✗ | ✗ | ✗ | 64.4 | 71.8 | 67.9 | 50.1 | 71.7 | 59.1 | **67.9** | 72.0 | 69.8 | 66.3 | 68.2 | 67.0 | 62.2 | 70.9 | 65.9 |
| 2 | ✓ | ✗ | ✗ | 66.9 | 72.2 | 69.5 | 52.9 | 74.8 | 62.0 | 65.1 | 78.3 | 71.1 | 66.5 | 74.1 | 70.1 | 62.9 | 74.9 | 68.2 |
| 3 | ✗ | ✓ | ✗ | **69.5** | 68.6 | 69.1 | 51.6 | 77.8 | 62.1 | 66.3 | 75.1 | 70.4 | **67.6** | 73.3 | 70.4 | 63.7 | 73.7 | 68.0 |
| 4 | ✗ | ✗ | ✓ | 63.3 | 74.3 | 68.4 | 50.3 | 77.2 | 60.9 | 64.9 | 76.4 | 70.2 | 66.0 | 76.2 | 70.7 | 61.1 | 76.0 | 67.5 |
| 5 | ✗ | ✓ | ✓ | 69.0 | 74.4 | 71.6 | 50.2 | **80.1** | 61.7 | 67.2 | 78.4 | 72.4 | 62.7 | **79.8** | 70.2 | 62.3 | **78.2** | 69.0 |
| 6 | ✓ | ✓ | ✓ | 68.3 | **79.9** | **73.7** | **54.6** | 74.8 | **63.1** | 66.9 | **80.2** | **73.0** | 66.1 | 77.3 | **71.3** | **64.0** | **78.2** | **70.3** |

Table 3. Ablation studiy results (%) on Office-Home with four different sub-tasks. The base and novel indicate the base-class and novel-class alignment heads in DCA, respectively. Row 1 indicates the reproduced baseline results with consistent implementation [3, 29].

| $\lambda_1$ | $\lambda_2$ | Office-Home | | | Image-CLEF | | |
|---|---|---|---|---|---|---|---|
| | | OS* | UNK* | **HOS** | OS* | UNK* | **HOS** |
| 0.5 | 1.0 | **66.2** | 73.1 | 69.5 | **78.6** | 83.6 | 80.9 |
| 1.0 | 1.0 | 65.6 | **76.7** | **70.7** | 78.2 | 85.6 | **81.4** |
| 2.0 | 1.0 | 64.9 | 76.2 | 70.1 | 76.1 | **87.2** | 81.2 |
| 1.0 | 0.5 | 64.2 | 75.3 | 69.3 | 77.2 | 81.6 | 79.0 |
| 1.0 | 1.0 | **65.6** | **76.7** | **70.7** | **78.2** | **85.6** | **81.4** |
| 1.0 | 2.0 | 64.9 | 75.1 | 69.6 | 78.0 | 84.9 | 81.3 |

Table 4. Sensitivity analysis on two benchmarks in terms of the loss weight terms $\lambda_{1/2}$. The default setting is $\lambda_{1/2} = 1.0$.

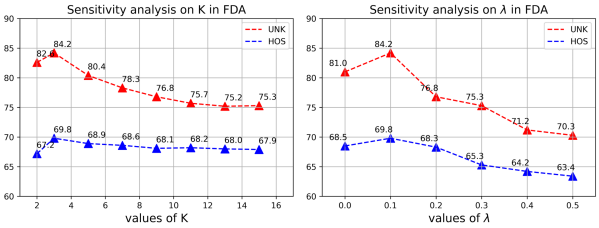


Figure 5. Sensitivity analysis of the hyperparameter, including (a) the K in discovering novel regions, and (b) the gradient scaling factor $\lambda$ in implementing the unbiased learning.

Liner-BN-LeakyReLU blocks and a binary domain classifier. Our model is trained with the Stochastic Gradient Descent optimizer with a 0.001 learning rate, 32 batch-size, a momentum of 0.9, and the most 100 epochs. The $\lambda$ and K in FDA are empirically set to 0.2 and 5, respectively.

## 5.2. Benchmark Comparison

**Office-Home.** We report the comparison results of Office-Home in Table 1 (top). The proposed method achieves the best average UNK (76.7%) and HOS (70.7%) over all 12 tasks, outperforming ROS [3], OSBP [29], and OSLPP [35] with 4.3%, 10.4 % and 5.0% UNK and 4.5%, 6.0% and 3.7% HOS, respectively. Compared with the state-of-the-art OSDA work OSLPP [35], our method comprehensively surpasses it with 1.8% OS*, 5.0% UNK, and 3.7% HOS, respectively. Moreover, ANNA achieves the best results in all 12 sub-tasks for the HOS comparison and 10 of 12 tasks for the UNK comparison, verifying the effect of our method.

**Image-CLEF.** The comparison is shown in Table 1 (bottom). In this real-world benchmark, we observe that ANNA achieves the best 81.4% HOS, which outperforms the other OSDA counterparts by a large margin, *e.g.*, yielding 8.3% gains than ROS [3], 9.9 % than OSBP [29] and 8.1% than DAOD [7], which verify our great potential for more complex real-world scenes. Moreover, our method achieves the best 85.6% UNK and a comparable 77.3% OS*, showing the adequate debiasing effect of the proposed method. Among all 12 sub-settings, ANNA achieves the best HOS in 10 of 12 tasks and the best UNK in 9 of 12 tasks, verifying the effectiveness of our method.

**Office-31.** Comparison results on Office-31 are shown in Table 2. ANNA gives the best average UNK (90.0%) and HOS (88.6%) evaluated over six tasks, which verifies the effectiveness of our method. Specifically, our method outperforms OSBP [29], ROS [3], and OSLPP [35] with 4.9%, 2.7%, and 1.2% HOS, and surpass them 9.6%, 4.2% and 4.4% in UNK comparison, demonstrating the robustness of our unbiased OSDA framework.

## 5.3. Ablation Study

As shown in Table 3, we conduct detailed ablation studies on Office-Home with four different settings and summarize the following observations. 1) Compared with the baseline model (65.9% HOS), introducing FDA (row 2, 68.2% HOS) and full DCA (row 5, 69.0% HOS) both give significant performance gains, verifying their individual effects. 2) Introducing FDA and DCA together (row 6) yields a further gain with the best 64.0% OS*, 78.2% UNK, and 70.3% HOS, showing the mutual benefits of the two modules. 3) As for DCA, we observe that introducing novel-

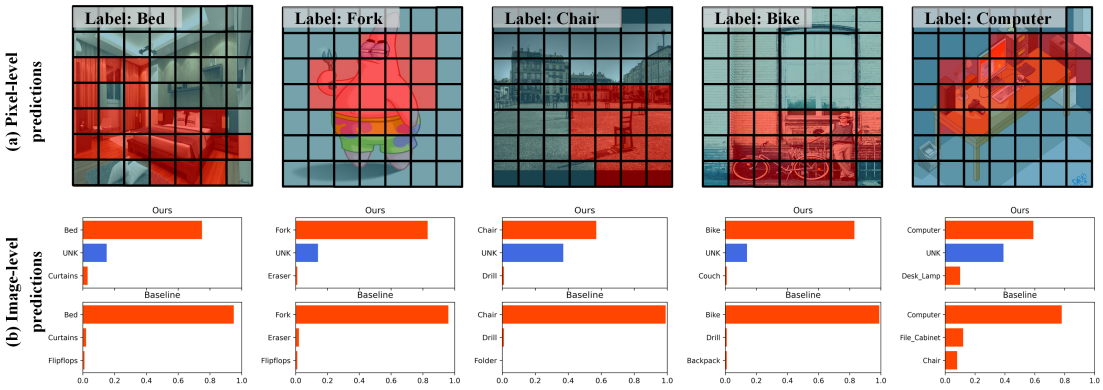


Figure 6. Visualization of (a) pixel-level and (b) image-level (Top-3 categories) predictions on the Office-Home benchmark. The color of red and blue indicates base-class and novel-class (UNK) predictions, respectively.

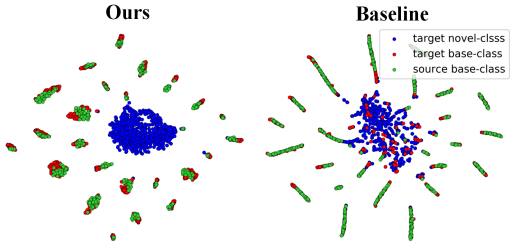


Figure 7. T-SNE feature comparison between the proposed ANNA (left) and baseline model (right) with Ar→Cl task.

class (row 4) alignment gives significant UNK gains (76.0% UNK) compared with baseline (70.9% UNK), and implementing the base-class alignment (row 3) improves OS* from 62.2% to 63.7%, showing the orthogonal effects of the decoupled two heads in the proposed DCA.

### 5.4. Sensitivity Analysis

**Loss Weight Terms.** We analyze the two loss terms $\lambda_{1/2}$ in Table 4. As for $\lambda_1$, compared with the default $\lambda_1 = 1.0$, setting a smaller ($\lambda_1 = 0.5$) value slightly improves OS* but harms UNK and HOS significantly, due to the sub-optimal debiasing effect. Increasing $\lambda_1$ from 1.0 to 2.0 gives a HOS decline from 70.7% to 70.1% (Office-Home) and from 81.4% to 81.2% (Image-CLEF), compared with $\lambda_1 = 1.0$. Moreover, we observe a slight decline in all evaluation metrics with a larger $\lambda_2 = 2.0$ and smaller $\lambda_2 = 0.5$.

**Discovering Novel Regions.** As shown in Figure 5(a), we explore the effect of K in discovering novel regions ($X_n$) with Ar→Cl task. With the increase of K, the number of discovered novel regions $X_n$ will decrease due to the more strict region selection constraint. We could observe a negative effect on novel-class accuracy (UNK) and HOS with a too-large K, verifying the effectiveness of the FDA.

**Unbiased Learning.** We analyze the effect of the gradient scaling factor $\lambda$ during optimizing $\mathcal{L}_{FDA}$, as shown in Figure 5(b). Setting a too-large value negatively affects both UNK and HOS. The reason may be that the dense pixel-level supervision can easily mislead the feature extractor to overfit partial low-quality backgrounds, which is risky for novel-class pattern learning with UNK decline.

### 5.5. Qualitative Results

**Output-level Analysis.** We visualize the (a) pixel-level and (b) image-level results and illustrate Top-3 predictions for each image, as shown in Figure 6. We observe that ANNA discovers novel regions and recognizes them as *unknown* through a per-pixel introspection (Figure 6(a)). Further, the novel regions are used to guide the image-level recognition without bias, yielding reasonable UNK scores (blue bars) in Figure 6(b). Moreover, as shown in $1^{st}$ image, we find that the score of *curtain* ranks in the second position ahead of other non-correlated categories, which verifies our practical design of eliminating $X_{ub}$ (Sec. 4.1).

**Feature-level Analysis.** As shown in Figure 7, we conduct a T-SNE feature comparison with the baseline. Our method is able to separate the *target novel-class* and *target base-class* more completely and thoroughly, benefiting a better base/novel-class decision boundary. Moreover, our method generates better embedding space in adapting *target base-class* and *source base-class* with different feature *clusters*, which is more discriminative and informative.

## 6. Conclusion

We observe the ever-overlooked bias in OSDA and propose a novel and theoretically grounded framework, Adjustment and Alignment (ANNA), to address it. ANNA adopts a Front-Door Adjustment to overcome the biased model learning in the source domain, which discovers the novel-class regions and grounds the causal debiasing theory with an unbiased learning loss. Besides, it leverages a Decoupled Causal Alignment (DCA) module to unbiasedly transfer the model to the target domain, disentangling the base/novel-class regions and aligning the decoupled conditional distribution. Extensive experiments on three standard benchmarks verify its state-of-the-art performance.

## References


[1] Image-clef. http://imageclef.org/2014/adaptation/, 2014. 6

[2] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009. 3

[3] 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. 1, 3, 4, 5, 6, 7

[4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 2018. 2

[5] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *CVPR*, pages 12455–12464, 2020. 1, 2, 5

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[7] Zhen Fang, Jie Lu, Feng Liu, Junyu Xuan, and Guangquan Zhang. Open set domain adaptation: Theoretical bound and algorithm. *TNNLS*, 32(10):4309–4322, 2021. 6, 7

[8] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015. 1, 2, 6

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 3, 6

[10] JoonHo Jang, Byeonghu Na, DongHyeok Shin, Mingi Ji, Kyungwoo Song, and Il-Chul Moon. Unknown-aware domain adversarial learning for open-set domain adaptation. In *Neurips*, 2022. 3

[11] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*, pages 4893–4902, 2019. 1, 2

[12] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *CVPR*, pages 10285–10295, 2019. 2

[13] Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. Maximum density divergence for domain adaptation. *PAMI*, 2020. 1, 2

[14] Wuyang Li, Xinyu Liu, Xiwen Yao, and Yixuan Yuan. Scan: Cross domain object detection with semantic conditioned adaptation. In *AAAI*, volume 36, pages 1421–1428, 2022. 2

[15] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, pages 5291–5300, June 2022. 2

[16] Xin Li, Zhizheng Zhang, Guoqiang Wei, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Confounder identification-free causal visual feature learning. *arXiv preprint arXiv:2111.13420*, 2021. 3

[17] 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. 1, 3, 4, 5, 6, 7

[18] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. *Neurips*, 34:22968–22981, 2021. 2

[19] Xinyu Liu, Wuyang Li, Qiushi Yang, Baopu Li, and Yixuan Yuan. Towards robust adaptive object detection under noisy annotations. In *CVPR*, pages 14207–14216, June 2022. 2

[20] Yadan Luo, Zijian Wang, Zi Huang, and Mahsa Baktashmotlagh. Progressive graph learning for open-set domain adaptation. In *ICML*, pages 6468–6478. PMLR, 2020. 1, 3, 5, 6

[21] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*, pages 415–430. Springer, 2020. 2

[22] Pedro R Mendes Júnior, Roberto M De Souza, Rafael de O Werneck, Bernardo V Stein, Daniel V Pazinato, Waldir R de Almeida, Otávio AB Penatti, Ricardo da S Torres, and Anderson Rocha. Nearest neighbors distance ratio open-set classifier. *Machine Learning*, 106(3):359–386, 2017. 6

[23] 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. 3

[24] Yulei Niu and Hanwang Zhang. Introspective distillation for robust question answering. *NeurIPS*, 34, 2021. 3

[25] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *ICCV*, pages 754–763, 2017. 1, 5

[26] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018. 2, 3

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 4

[28] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226. Springer, 2010. 6

[29] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *ECCV*, pages 153–168, 2018. 1, 2, 3, 4, 5, 6, 7

[30] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007. 3

[31] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *CVPR*, pages 8725–8735, 2020. 2

[32] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020. 3

[33] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*. 4

[34] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017. 4, 6

[35] Qian Wang, Fanlin Meng, and Toby P Breckon. Progressively select and reject pseudo-labelled samples for open-set domain adaptation. *AAAI*, 2022. 1, 3, 5, 6, 7

[36] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *CVPR*, pages 10760–10770, 2020. 1, 2, 3, 5

[37] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *ICCV*, pages 3091–3100, 2021. 3

[38] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. In *AAAI*, volume 33, pages 5345–5352, 2019. 2, 5

[39] Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *ICLR*, 2022. 3

[40] Hang Xu, ChenHan Jiang, Xiaodan Liang, Liang Lin, and Zhenguo Li. Reasoning-rcnn: Unifying adaptive global reasoning into large-scale object detection. In *CVPR*, pages 6419–6428, 2019. 4

[41] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Universal domain adaptation. In *CVPR*, June 2019. 6, 7

[42] Zhongqi Yue, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Transporting causal mechanisms for unsupervised domain adaptation. In *ICCV*, pages 8599–8608, 2021. 2, 3

[43] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *CVPR*, pages 15404–15414, 2021. 2, 3

[44] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *NeurIPS*, 2020. 1, 2, 3

[45] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *NeurIPS*, 33, 2020. 1, 2, 3, 5

[46] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *CVPR*, pages 5031–5040, 2019. 2