Activate and Reject: Towards Safe Domain Generalization under Category Shift

Chaoqi Chen[1*], Luyao Tang[2*], Leitian Tao[3], Hong-Yu Zhou[1], Yue Huang[2], Xiaoguang Han[4†], Yizhou Yu[1†]
[1] The University of Hong Kong [2] Xiamen University
[3] University of Wisconsin - Madison [4] The Chinese University of Hong Kong (Shenzhen)
cqchen1994@gmail.com, lytang@stu.xmu.edu.cn, taoleitian@gmail.com, whuzhouhongyu@gmail.com
yhuang2010@xmu.edu.cn, hanxiaoguang@cuhk.edu.cn, yizhouy@acm.org

Abstract

Albeit the notable performance on in-domain test points, it is non-trivial for deep neural networks to attain satisfactory accuracy when deploying in the open world, where novel domains and object classes often occur. In this paper, we study a practical problem of Domain Generalization under Category Shift (DGCS), which aims to simultaneously detect unknown-class samples and classify known-class samples in the target domains. Compared to prior DG works, we face two new challenges: 1) how to learn the concept of "unknown" during training with only source known-class samples, and 2) how to adapt the source-trained model to unseen environments for safe model deployment. To this end, we propose a novel Activate and Reject (`ART`) *framework to reshape the model's decision boundary to accommodate unknown classes and conduct post hoc modification to further discriminate known and unknown classes using unlabeled test data. Specifically, during training, we promote the response to the unknown by optimizing the unknown probability and then smoothing the overall output to mitigate the overconfidence issue. At test time, we introduce a step-wise online adaptation method that predicts the label by virtue of the cross-domain nearest neighbor and class prototype information without updating the network's parameters or using threshold-based mechanisms. Experiments reveal that* `ART` *consistently improves the generalization capability of deep networks on different vision tasks. For image classification,* `ART` *improves the H-score by 6.1% on average compared to the previous best method. For object detection and semantic segmentation, we establish new benchmarks and achieve competitive performance.*

1. Introduction

Deep neural networks have achieved unprecedented success in a myriad of vision tasks over the past decade. Despite the promise, a well-trained model deployed in the open and ever-changing world often struggles to deal with the domain shifts—the training and testing data do not follow the independent and identically distributed (i.i.d) assumption, and therefore deteriorates its safety and reliability in many safety-critical applications, such as autonomous driving and computer-aided disease diagnosis. This gives rise to the importance of Domain Generalization (DG) [101, 83], *a.k.a.* out-of-distribution (OOD) generalization, which aims at generalizing predictive models trained on multiple (or a single) source domains to unseen target distributions.

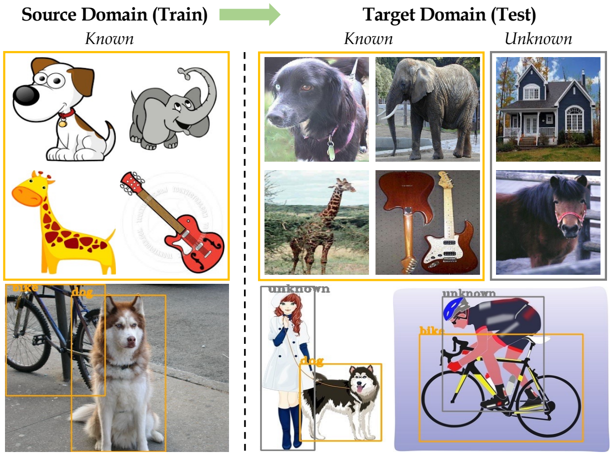


Figure 1. DGCS in image classification and object detection tasks.

In order to unearth domain-agnostic knowledge and alleviate domain-specific components, a plethora of DG algorithms have been proposed, spanning invariant risk minimization [2, 1], augmentation [81, 87, 105, 9], feature disentanglement [63, 49, 93], meta-learning [43, 44, 21], to name a few. Among them, a common assumption is that the label spaces of source and target domains are identical, which may not always hold in practice. Suppose that we wished to deploy modern vision systems to recognize objects in an autonomous vehicle. When only the environment (*e.g.,* weather and illumination) and appearance (*e.g.,* size and viewpoint) of previously seen objects can change, principled approaches are capable of correcting for the po-

[*] First two authors contributed equally.
[†] Corresponding authors.

tential shifts on the fly. But what if the sudden arrival of new objects in an ever-changing world? Most existing DG methods will break and may even result in catastrophe, raising strong concerns about model reliability. Although several prior arts [71, 106] have explored the open DG scenarios, the "adaptivity gap" [22] between training and test distributions still hinders safe deployment of source-learned models [30].

To this premise, we challenge the status quo by raising an open question: *can deep models learn what they don't know during training and subsequently adapt to novel environments at test-time for safe model deployment?* Thus, we consider a more realistic scenario namely Domain Generalization under Category Shift (DGCS) (see Fig. 1), wherein the source-trained model is expected to simultaneously detect unknown-class samples and categorize known-class samples under the presence of domain shifts. The core challenges are: *(i)* no unknown-class data is available in training and *(ii)* the mixture of domain and label shifts during test time. In this paper, we present a simple yet effective framework—**A**ctivate and **R**ejec**T** (dubbed `ART`), which reshapes the model's decision boundary to accommodate unknown classes and adjusts the final prediction to reconcile the intrinsic tension between domain and label shifts. `ART` encapsulates two key components: *(i)* Unknown-aware Gradient Diffusion (UGD) to make the classifier give response to unknown dimension and smooth the decision boundary to mitigate overconfidence; *(ii)* Test-time Unknown Rejection (TUR) to conduct *post hoc* modification to the learned classifier's final predictions, making the decision boundaries of different classes closer to the well-behaved case.

Specifically, the logit of unknown class is activated by minimizing the negative log-likelihood regarding unknown probability. However, we find that the learned probability will be suppressed due to the overconfidence *w.r.t.* known classes. Thus, we introduce a smoothed cross-entropy loss to promote the response to the unknown by adding the penalty on the L_2 norm of the logits and using a temperature scaling parameter, where the former mitigates the excessive increase of the logit norm while the latter magnifies the effect of logit penalty. Due to the unavailability of real target data in training, the source-trained decision boundaries between known and unknown classes may still be ambiguous. Therefore, TUR refines the source-trained classifier using unlabeled test data in an online adaptation manner. To be specific, TUR first determines if the input belongs to known classes or not via a cross-domain nearest neighbor search, based on prototype information and cyclic consistent constraint; otherwise, the prediction will be made by a parallel module that measures the input's similarity with a set of dynamically-updated target prototypes. TUR is training-free (no backward passes) and does not rely on threshold-based criteria nor impose any distributional assumptions.

Our key contributions are summarized as follows:

- We study a challenging DG problem (DGCS) and propose a principled framework (`ART`) to jointly consider domain shift, label shift, and adaptivity gap.
- We propose an unknown-aware training objective to activate the unknown's logit and alleviate the overconfidence issue, and an online adaptation strategy to perform post hoc modification to the learned classifier's prediction at test-time without additional tuning.
- Extensive experiments show that `ART` achieves superior performance on a wide range of tasks including image classification, object detection, and semantic segmentation. In particular, on four image classification benchmarks (PACS, Office-Home, Office-31, and Digits), `ART` improves the H-score by 6.1% on average compared to the previous best method.

2. Related Works

Domain Generalization (DG). The objective of DG is to learn representations that are independent of domain-specific factors and thus can extrapolate well to unseen test distributions. This is typically achieved by invariant learning and robust learning. Current approaches can be broadly categorized into feature matching [45, 54, 107, 12], decomposition [66, 63, 17, 53, 72, 49, 93, 100], augmentation [80, 102, 103, 87, 56, 105, 86, 96, 13, 6], and meta-learning-based [43, 46, 44, 21, 15] approaches. To adapt to complex real-world applications, very recently, several works [71, 106, 91] consider the existence of both known and unknown classes in new DG settings, such as open DG [71] open-set DG (OSDG) [106]. Shu *et al.* [71] assume that both source and target domains have different label spaces and introduce novel augmentation strategies to augment domains on both feature- and label-level. Zhu *et al.* [106] generate auxiliary samples via an adversarial data augmentation strategy [81] and enhance unknown class identification with multi-binary classifiers. Yang *et al.* [91] introduce an additional CE loss based on the assumption that any non-ground-truth category can be viewed as unknown categories. However, these works rely on additional training modules and heuristic thresholding mechanism [106] or impose a strong distributional assumption of the feature space regarding known and unknown data [91]. In addition, Dubey *et al.* [22] reveal that there will always be an "adaptivity gap" when applying the source-learn model to target domains without further adaptation. How to endow the source model with the capability of identifying unseen open classes and safely adapting the learned classifier to unlabeled test samples is yet to be thoroughly studied.

Domain Adaptation (DA). DA [59, 52, 25, 10, 8] aims to improve the performance of the learned model on the target domain using labeled source data and unlabeled target

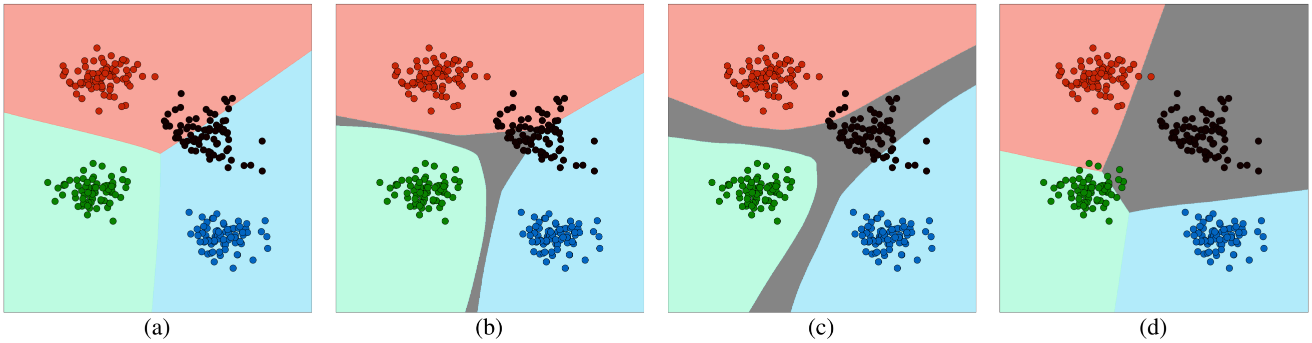


Figure 2. Toy example illustrating the decision boundaries learned by different methods. We generate isotropic Gaussian blobs with 4 classes. Red, green, and blue points indicate the known-class samples. Black points denote unknown-class samples, which *are unavailable during training*. **(a)** Train with standard CE loss, *i.e.,* vanilla ($|\mathcal{C}_s|$+1)-way classifier in DGCS. **(b)** Train with our unknown activation loss $\mathcal{L}_{\text{UA}}$. **(c)** Train with full UGD loss $\mathcal{L}_{\text{UGD}}$. **(d)** The result of ART (UGD + TUR). This figure is best seen in color.

data. In addition to the close-set setting, many new and practical DA paradigms have been proposed, such as partial [94, 5], open-set [60, 69, 38, 4, 7, 50], universal [92, 68], and source-free [85, 89, 20, 97, 90]. In particular, open-set DA (OSDA) and source-free DA (SFDA) are closely related to the problem explored in this paper.

Test-Time Adaptation (TTA). For DG, due to the inaccessibility of target data during training, it is natural to solve the adaptivity gap [22] with TTA strategies. Adaptive methods [47, 76, 82, 35, 61, 11, 95, 14] have been proposed to refine the matching process between target test data and source-trained models in an online manner, *i.e.,* all test data can be accessed only once. Tent [82] proposes to reduce the entropy of model's predictions on test data via entropy minimization. T3A [35] introduces a training-free approach by classifying each test sample based on its distance to a dynamically-updated support set. Despite the promising results on closed-set classes, these approaches fail to deal with open-set samples and thus lead to semantic mismatching.

Out-of-Distribution Detection (OD). A separate line of work studies the problem of OD [88], which aims to identify novel examples that the network has not been exposed to at the training phase. Mainstream OD methods are devoted to design OOD scoring functions, *e.g.,* confidence-based approaches [3, 31, 32], distance-based score [41, 70, 75], and energy-based score [51, 73]. The main difference between OD and our problem is that the former is a binary classification problem and does not account for the domain and label shifts between training and test data at the same time.

Discussion. We provide a comparison of the problem settings among different methods in Tab. 1. OSDA and SFDA optimize offline with target data and specific learning objectives, while ART only adjusts the classifier in an online manner. TTA usually needs to update the trained model's parameters (*e.g.* entropy minimization [82, 95]) and a batch of data, while our TUR is fully training-free and can be performed on single test samples. These promising properties make the proposed approach more suitable for DG. Compared to OSDG, our setting allows training-free test-time adjustment for adapting source-trained models to novel environments, largely mitigating the potential adaptivity gap.

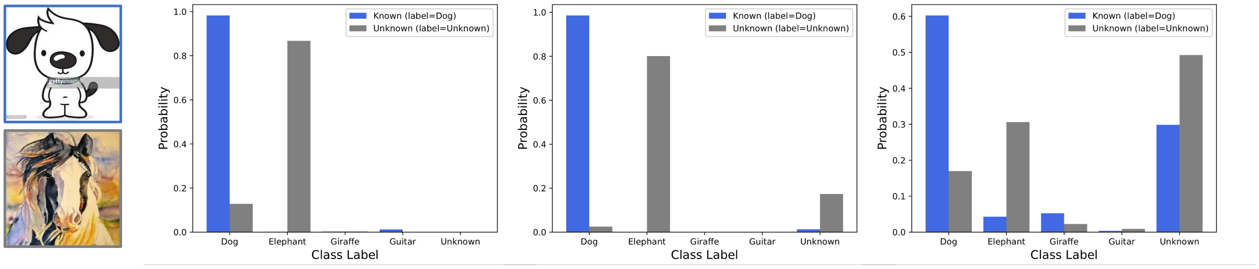
Table 1. Comparison of different problem settings. (X_s, Y_s) and X_t are the labeled source and unlabeled target data respectively. Fine-tune means to update the model's parameters. Adjustment means making post-hoc modifications to the model's predictions.

Problem Setting	**Training**	**Test-time**			
	Data	Domain Shift	Open Class	Fine-tune	Adjustment
OD [31, 74]	X_s, Y_s	✗	✓	✗	✗
OSDA [69, 4]	X_s, Y_s, X_t	✓	✓	✗	✗
SFDA [92, 97]	X_s, Y_s, X_t	✓	✗	✓	✗
TTA [76, 82, 95]	X_s, Y_s	✓	✗	✓	✓
OSDG [106, 91]	X_s, Y_s	✓	✓	✗	✗
Ours	X_s, Y_s	✓	✓	✗	✓

3. Methodology

3.1. Preliminary and Motivation

Notation. In DGCS, we have a single source domain $\mathcal{D}_s = \{(x_s^i, y_s^i)\}_{i=1}^{n_s}$ of n_s labeled samples and multiple (or a single) unseen target domains $\mathcal{D}_t = \{\mathcal{D}_t^1, ..., \mathcal{D}_t^M\}$, where $M \geq 1$ and $\mathcal{D}_t^m = \{(x_t^j, y_t^j)\}_{j=1}^{n_t^m}$. $\mathcal{D}_s$ and $\mathcal{D}_t$ are sampled from probability distributions $p_s(x, y)$ and $p_t(x, y)$ respectively. DGCS jointly considers two distribution shifts: *(i)* class-conditional shift where $p_s(y|x) \neq p_t(y|x)$, and *(ii)* label shift where $p_s(y) \neq p_t(y)$. Specifically, assume that $\mathcal{C}_s$ and $\mathcal{C}_t$ are the source and target class sets, respectively. DGCS dictates $\mathcal{C}_s \subset \mathcal{C}_t$ and $\mathcal{C}_t^u = \mathcal{C}_t \setminus \mathcal{C}_s$ is called *unknown* classes. Note we take all unknown classes as a whole even though there can be multiple classes. The objective of



Input Image (a) Vanilla Classifier (b) Unknown Activation (c) Full UGD Loss

Figure 3. The softmax outputs of different training methods regarding two input images on the PACS [42] benchmark. Source Domain: *Cartoon*, Target Domain: *Art*. Known: *dog* from Domain *Cartoon*, Unknown: *horse* from Domain *Art*.

DGCS is to train a model on $\mathcal{D}_s$ to classify all target instances from $\mathcal{D}_t$ into $|\mathcal{C}_s|+1$ classes.

Motivation. Before formally introducing technical details, we discuss the motivation of our method using toy data. Since the decision boundaries are learned by known classes only, the unknown target samples tend to lie out of the support of source training data (*i.e.,* low-density regions [27]) and are ambiguous for the decision boundaries. On the other hand, as shown in Fig. 3(a), deep neural networks trained with the standard softmax Cross-Entropy (CE) loss tend to give overconfident predictions even when the test input differs from the training distribution [58]. Motivate by this, our goal is to explicitly create a support region for unknown target samples. A native choice is the low-density regions with respect to the source-trained classifier.

To empirically verify our intuitions, we use *scikit-learn* [62] to generate samples (3 known classes and 1 unknown class) and show the comparison in Fig. 2. From the figure, we have the following observations. (1) Simply training a ($|\mathcal{C}_s|$+1)-way classifier cannot improve the discrimination of unknown class. (2) Forcefully increasing the softmax probability in the unknown dimension creates an additional support region. However, due to the overconfidence issue regarding known classes, the response to the unknown (reflected by the size of the region) is still limited. (3) To increase the response to unknown class, we penalize the prediction confidence *w.r.t.* known classes, *i.e.,* making the known-class data closer to their decision boundaries. (4) Although the reshaped decision boundaries are able to accommodate unknown-class data, the boundaries between known and unknown classes are less discriminative as we do not have access to real unknown data, *i.e.,* the unknown samples do not necessarily lie in the support of the created region since the above operations only encourage it far away from the support of known classes. Thus, we dynamically adjust the learned boundaries using unlabeled test data.

Grounded on these insights, we propose a novel Activate and Reject (ART) approach. Specifically, ART encompasses two innovative components: 1) Unknown-aware Gradient Diffusion (UGD) to diffuse the gradient to the unknown's logit with smoothing regularization; 2) Test-time Unknown Rejection (TUR) to conduct post hoc modification to the learned classifier's final prediction.

3.2. Unknown-aware Gradient Diffusion

As discussed in Sec. 3.1, deep classifiers trained with the standard softmax CE loss are susceptible to the notorious overconfidence issue. This problem becomes more sophisticated in the context of DGCS, wherein the learned decision boundary is highly biased towards source known-class samples. On the other hand, given only access to known-class data during the training phase, how to optimize the $|\mathcal{C}_s|$+1-way classifier is problematic (cf. Fig. 2(a)).

With this premise, we propose the UGD to solve the above issues at training phase from two perspectives, *i.e.,* unknown activation and output's smoothness. The former activates the unknown probability, while the latter mitigates the overconfidence issue. First of all, we need to train a ($|\mathcal{C}_s|$+1)-way classifier, where an additional dimension is introduced to discriminate unknown classes from known ones. Given $\mathbf{x}_s \in \mathcal{D}_s$ and a neural network $f(\mathbf{x};\theta)$ parameterized by θ, we define the standard CE loss as:

$$\mathcal{L}_{\mathrm{CE}}(f(\mathbf{x_s}), y_s) = -\log \frac{\exp(f_{y_s}(\mathbf{x_s}))}{\sum_{k\in|\mathcal{C}_s|+1} \exp(f_k(\mathbf{x_s}))}, \quad (1)$$

where $f(\mathbf{x_s}) \in \mathbb{R}^{|\mathcal{C}_s|+1}$ denotes the network's logit and $f_{y_s}(\mathbf{x_s})$ is the y_s-th element of $f(\mathbf{x_s})$ corresponding to the ground-truth label y_s.

Based on the ($|\mathcal{C}_s|$+1)-way classifier, we aim to activate the unknown's logit in the absence of real unknown-class samples. The key idea is to *increase the value of unknown probability without affecting the ground-truth classification.* For notation shorthand, we use $\boldsymbol{f}_k$ to represent the logit of k-th class and $\boldsymbol{f}_u$ for the unknown's logit. Since we have no supervision over the unknown, the value of $\boldsymbol{f}_u$ is negligible (cf. Fig. 3(a)). For a source sample $(\mathbf{x}_s, y_s) \in \mathcal{D}_s$, we forcefully increase the unknown probability by minimizing the negative log-likelihood,

$$\mathcal{L}_{\mathrm{UA}} = -\log \frac{\exp(\boldsymbol{f}_u)}{\sum_{k\in|\mathcal{C}_s|+1, k\neq y_s} \exp(\boldsymbol{f}_k)}, \quad (2)$$

This objective ensures that the unknown probability can give a response to any input sample regardless of its class label (cf. gray region in Fig. 2(b)). Since the learning process is always dominated by CE loss regarding the ground-truth category, Eq. (2) is tractable and will not hurt the known-class performance. However, the activated probability is relatively small (compared to the ground-truth category), which leads to an unsatisfactory accuracy for real unknown samples, especially for some hard samples (cf. Fig. 3(b)).

Next, we aim to enhance the response to unknown classes by increasing the smoothness of the network's output (cf. Fig. 2(c)). Formally, we impose two constraints to the standard CE loss: a temperature scaling parameter τ ($\tau > 1$) and a penalty on the L_2 norm of the logits. Thus, the proposed smoothed CE (SCE) loss $\mathcal{L}_{\text{SCE}}$ is defined as:

$$\mathcal{L}_{\text{SCE}} = -\log \frac{\exp(f_{y_s}(\mathbf{x_s})/\tau)}{\sum_{i \in |\mathcal{C}_s|+1} \exp(f_i(\mathbf{x_s})/\tau)} + \lambda \|f(\mathbf{x_s})\|_2, \tag{3}$$

where λ is set to 0.05 in all experiments.

Finally, the UGD loss is formulated as:

$$\mathcal{L}_{\text{UGD}} = \mathcal{L}_{\text{UA}} + \mathcal{L}_{\text{SCE}}. \tag{4}$$

As shown in Fig. 3(c), the proposed $\mathcal{L}_{\text{UGD}}$ not only reduces the overconfidence issue (smaller max-probability for known sample) but also significantly increases the unknown probability.

3.3. Test-Time Unknown Rejection

Although we have activated the network's logit about unknowns, there still exist two critical challenges that impede the safe and reliable deployment of our source-trained models on open-world data. First, a conservative (smaller max-probability) and smoothing (larger entropy) output on source data may not guarantee the category correspondence across domains and therefore may lead to semantic misalignment. Second, how to reject a sample as "unknown" lacks principled criterion considering that the unknown-class samples may distribute randomly in the embedding space. In this regard, previous open-set-oriented methods [68, 106] that typically rely on thresholding mechanisms (*e.g.* entropy value [106]) are heuristic and will be sensitive to the variations of domain disparity.

To solve the above issues, we introduce a simple and effective technique—TUR—to match unlabeled test data to the source-trained model in an online adaptation manner. Our key idea is to conduct *post hoc* modification to the learned classifier's final predictions, so as to bring the decision boundaries of different classes closer to the well-behaved case. TUR is *training-free* (*i.e.,* no backward passes) and does not impose any distributional assumptions.

Technically, we impose a cross-domain cycle-consistent constraint on the top of embedding space for identifying whether a test sample corresponds to any known classes or not. The cross-domain relationships are based on K-nearest neighbor (KNN) [36] to perform non-parametric density estimation, which is model-agnostic and easy to implement. Specifically, we decompose the source-trained model into a feature extractor g and a linear classifier f. Assume that the embedding of training data is $\mathbb{Z}_s = \{\boldsymbol{z}_s^1, \boldsymbol{z}_s^2, ..., \boldsymbol{z}_s^{n_s}\}$, where $\boldsymbol{z}_s^i$ is the L_2-normalized penultimate feature $\boldsymbol{z}_s^i = g(\mathbf{x}_s)/\|g(\mathbf{x}_s)\|_2$. Here, we do not require access to the original training samples since the embedding will be extracted in advance, and no need to update. Then, we define two sets of known-class prototypes on the top of penultimate layer, *i.e.,* $\{\mu_s^k\}_{k=1}^{|\mathcal{C}_s|}$ and $\{\mu_t^k\}_{k=1}^{|\mathcal{C}_s|}$, where μ_s^k is computed from $\mathbb{Z}_s$ (mean feature per class) and will be fixed at test time. μ_t^k is empty at the beginning.

For an test input $\mathbf{x}_t^j$ with its normalized feature vector $\boldsymbol{z}_t^j$, we compute its KNN in $\mathbb{Z}_s$, denoted by $\mathcal{N}_s(\boldsymbol{z}_t^j)$. The feature centroid of $\mathcal{N}_s(\boldsymbol{z}_t^j)$ is denoted by $\bar{\boldsymbol{z}}_s^j$. Next, we find the corresponding source class as,

$$k' = \underset{k' \in \{0,1,\cdots,|\mathcal{C}_s|\}}{\arg\max} \; sim(\bar{\boldsymbol{z}}_s^j, \mu_s^{k'}) \tag{5}$$

Here, we measure the cosine similarity between features as: $sim(\bar{\boldsymbol{z}}_s^j, \mu_s^{k'}) = \frac{(\bar{\boldsymbol{z}}_s^j)^T \mu_s^{k'}}{\|\bar{\boldsymbol{z}}_s^j\|_2 \|\mu_s^{k'}\|_2}$. In the same way, we search the target class k'' based on the similarity between $\bar{\boldsymbol{z}}_s^j$ and $\mu_t^{k''}$. If k' and k'' belongs to the same category, the sample $\boldsymbol{x}_t^j$ will be predicted as class k'' and we further update $\mu_t^{k''}$ in the following manner,

$$\mu_{t(I)}^{k''} = \phi \, \boldsymbol{z}_t^j + (1-\phi) \, \mu_{t(I)}^{k''}, \tag{6}$$

where $\mu_{t(I)}^{k''}$ denote the k''-th target prototype until time I and $\phi \in (0, 1)$ is a preset scalar and fixed to 0.3 in practice.

If k' and k'' belong to different categories, the prediction will be given by using a follow-up strategy. Specifically, a memory bank $\mathbb{M}_I = \{\mathbb{M}_I^1, \cdots, \mathbb{M}_I^{|\mathcal{C}_s|+1}\}$ is a set of target sample embedding until time I, which is initialized by the weight of linear classifier f. At time I, $\mathbb{M}_I$ is updated as:

$$\mathbb{M}_I^k = \begin{cases} \mathbb{M}_{I-1}^k \cup \boldsymbol{z}_t^j & \text{if } k' \neq k'' \text{ and } f(\boldsymbol{z}_t^j) = k, \\ \mathbb{M}_{I-1}^k & \text{otherwise,} \end{cases} \tag{7}$$

Similarly, we can build a new set of target class prototypes $\{\psi_t^k\}_{k=1}^{|\mathcal{C}_s|+1}$ based on samples from $\mathbb{M}_I$. Note that ψ_t^k will be constantly updated during test time. Then, we predict the class label (($|\mathcal{C}_s|$+1)-way) of $\mathbf{x}_t^j$ as follows,

$$\hat{k} = \underset{\hat{k} \in \{0,1,\cdots,|\mathcal{C}_s|+1\}}{\arg\max} \; sim(\bar{\boldsymbol{z}}_s^j, \psi_t^{\hat{k}}). \tag{8}$$

The decision boundaries between known and unknown classes are refined without backpropagation (cf. Fig. 2(d)).

Table 2. Accuracy (%) on four classification benchmarks (ResNet-18).

Regime	Method	PACS			Office-Home			Office-31			Digits			Average		
		acc_k	acc_u	hs	acc_k	acc_u	hs	acc_k	acc_u	hs	acc_k	acc_u	hs	acc_k	acc_u	hs
OSDA (upper bound)	OSBP [69]	40.6	49.5	44.6	47.1	66.9	55.3	75.8	84.3	77.7	35.6	70.6	40.5	49.8	67.8	54.5
	ROS [4]	35.6	66.4	46.4	50.8	77.5	60.8	71.7	80.0	75.6	20.1	48.6	34.9	47.7	68.1	54.4
OD	MSP [31]	38.9	62.5	46.4	52.7	75.6	62.0	49.7	89.2	63.8	17.2	87.1	28.8	39.6	78.6	50.3
	LogitNorm [84]	35.1	47.6	38.3	56.3	56.5	56.1	41.0	71.2	52.1	26.8	51.2	35.2	39.8	56.6	45.4
	DICE [74]	44.0	53.4	49.2	61.5	58.8	59.9	72.8	61.1	66.4	35.0	47.6	40.3	53.3	55.2	54.0
SFDA	SHOT [47]	51.2	34.9	40.8	52.5	32.4	44.3	84.8	60.2	70.4	27.4	20.3	23.3	54.0	37.0	44.7
	AaD [90]	45.1	40.0	42.0	59.4	58.7	58.9	70.1	85.3	76.9	25.6	26.9	26.2	50.1	52.7	51.0
TTA	TTT [76]	36.9	44.6	38.9	52.0	45.9	47.2	35.4	79.6	49.0	44.1	45.1	44.6	42.1	53.8	44.9
	Tent [82]	25.2	43.1	31.7	33.6	45.9	38.7	56.0	85.1	67.5	27.2	41.1	32.7	35.5	53.8	42.7
	MEMO [95]	37.9	52.3	44.5	49.0	55.6	52.1	59.8	72.7	65.6	21.7	56.1	31.3	42.1	59.2	48.4
OSDG	ERM [78]	52.3	27.0	36.1	66.9	23.7	34.3	85.1	27.0	40.7	56.4	13.0	18.0	65.2	22.7	32.3
	ADA [81]	54.2	30.9	36.4	67.9	25.4	36.2	85.6	25.2	38.7	57.2	15.1	20.1	66.2	24.2	32.9
	ADA+CM [106]	56.4	45.6	43.0	65.0	40.4	48.5	83.0	34.5	48.5	49.2	52.1	39.9	63.4	43.2	45.0
	MEADA [98]	54.1	31.4	36.2	67.6	25.7	36.4	85.8	25.1	38.6	57.6	29.8	30.4	66.3	28.0	35.4
	MEADA+CM [106]	54.3	46.6	42.7	64.9	40.5	49.6	82.8	41.1	54.7	52.3	46.1	38.7	63.6	43.6	46.4
	One Ring-S [91]	43.7	49.4	41.5	56.9	69.0	62.3	67.3	77.0	71.3	33.2	51.3	40.3	50.3	61.7	53.9
	ART w/o TUR	47.0	51.3	48.1	58.8	69.8	63.7	70.7	65.9	68.2	29.7	65.7	40.9	51.6	63.2	55.2
DGCS	ART (full)	43.7	65.9	**52.3**	64.3	65.3	**64.8**	82.1	75.2	**78.5**	34.3	63.8	**44.6**	56.1	67.6	**60.1**

Table 3. Performance of ART on object detection benchmarks.

Method	*Pascal VOC→Clipart*					*Pascal VOC→Watercolor*					*Pascal VOC→Comic*				
	WI$_\downarrow$	AOSE$_\downarrow$	mAP$_{\mathcal{K}\uparrow}$	AP$_{\mathcal{U}\uparrow}$	$hs_\uparrow$	WI$_\downarrow$	AOSE$_\downarrow$	mAP$_{\mathcal{K}\uparrow}$	AP$_{\mathcal{U}\uparrow}$	$hs_\uparrow$	WI$_\downarrow$	AOSE$_\downarrow$	mAP$_{\mathcal{K}\uparrow}$	AP$_{\mathcal{U}\uparrow}$	$hs_\uparrow$
ORE [37]	17.3	876	**37.7**	3.0	5.6	28.4	3216	19.8	13.5	16.1	23.1	2242	7.3	3.0	4.3
OpenDet [28]	14.2	**300**	32.7	6.7	11.1	**14.9**	1944	19.2	**19.3**	19.2	15.2	744	**7.5**	3.1	4.4
ART (full)	**11.7**	317	35.8	**10.2**	**15.9**	19.7	**944**	20.8	15.2	17.6	**13.2**	**596**	7.2	**9.1**	**8.0**
w/o $\mathcal{L}_{\text{UA}}$	16.3	1363	35.4	6.0	10.3	29.4	3924	18.6	14.1	16.0	25.0	2826	6.4	2.2	3.3
w/o $\mathcal{L}_{\text{SCE}}$	14.6	426	34.7	3.2	5.9	24.8	1104	**21.4**	19.1	**20.2**	24.7	1372	6.3	3.5	4.5
w/o TUR	14.9	444	34.5	4.9	8.6	23.3	1398	21.1	17.6	19.2	15.0	784	7.3	4.6	5.6

Table 4. Performance of ART on semantic segmentation benchmark, *i.e.,* from **GTA5** (synthetic) to **Cityscapes** (real).

Method	mAcc	mIOU	acc_u	hs
ERM [78]	64.9	48.2	27.6	39.4
One Ring-S [91]	55.7	41.0	72.5	61.9
ART (full)	57.1	43.3	73.2	**63.1**
w/o UGD	64.2	46.6	41.6	50.2
w/o TUR	54.7	42.6	78.5	62.6

4. Experiments

4.1. Generalization in Image Classification

Dataset. We evaluate our ART on four standard DG benchmarks. **PACS** [42], which has dramatic differences in terms of image styles, contains 9,991 images of seven object classes from four domains, *i.e., Photo*, *Art Painting*, *Cartoon*, and *Sketch*. 4 classes (dog, elephant, giraffe, and guitar) are adopted as $\mathcal{C}_s$ and the remaining 3 classes are used as $\mathcal{C}_u^t$. **Office-Home** [79], which is collected from office and home environments, has 15,500 images of 65 classes from four domains, *i.e., Artistic*, *Clipart*, *Product*, and *Real World*. The domain shifts stem from the variations of viewpoint and image style. In alphabetic order, the first 15 classes are selected as $\mathcal{C}_s$ and the remaining 50 classes are used as $\mathcal{C}_u^t$. **Office-31** [67] has 31 classes collected from three domains: *Amazon*, *DSLR*, and *Webcam*. The 10 classes shared by Office-31 and Caltech-256 [26] are adopted as $\mathcal{C}_s$. In alphabetical order, the last 11 classes along with $\mathcal{C}_s$ form $\mathcal{C}_u^t$. **Digits**, which differs in the background, style, and color, contains four handwritten digit domains including *MNIST* [40], *MNIST-M* [25], *SVHN* [57], *USPS* [33], and *SYN* [25]. *MNIST* is used as the source domain and the other datasets are viewed as target domains. $\mathcal{C}_s$ includes numbers from 0 to 4.

Evaluation Protocols. Following [4, 106, 91], we adopt H-score (hs) [24] as the main evaluation metric. hs harmonizes the importance of known and unknown classes by requiring that known and unknown class accuracy should be both high and balanced. The known class accuracy (acc_k) and unknown class accuracy (acc_u) are also provided.

Implementation Details. We conduct experiments based on Dassl [104], including data preparation, model training, and model selection. For PACS, Office-Home, and Office-31, we use ResNet-18 [29] pre-trained on the ImageNet as the backbone network. We use the ConvNet [39] with architecture *conv-pool-conv-pool-fc-fc-softmax* for Digits. The networks are trained using SGD with momentum of 0.9 for 100 epochs. The batch size is set to 16.

Baselines. Given the contact points with other problem settings, we compare ART with five types of state-of-the-art methods. (1) **OSDG** [106, 91] is the most related baseline. *When TUR is removed, the proposed ART becomes a standard OSDG method.* (2) **OSDA** [69, 4] jointly utilizes source and target data for training and thus can be viewed as an upper bound of our problem. (3) **OD** [84, 74] usually identifies unknown-class samples via scoring functions. (4) **SFDA** [47, 90] and **TTA** [47, 90] cannot deal with unknown-class samples directly. Therefore, we follow [106] that uses the entropy of softmax output as the normality score.

Results. The classification results on PACS and Office-Home, Office-31, and Digits benchmarks are reported in Tab. 2. ART substantially and consistently outperforms baseline methods on different benchmark datasets. For example, ART improves hs by 9.3% (PACS), 2.5% (Office-Home), 7.2% (Office-31), and 4.3% (Digits) compared to the previous best OSDG baselines. In particular, only using UGD could also substantially exceed state-of-the-art methods, *e.g.* DICE [74] and One Ring-S [91]. The results also reveal several interesting observations. (1) The performance of [91] is unstable across different benchmarks. For example, they outperform CM [106] by +12.7% and +16.6% on Office-Home and Office-31 but show inferior performance (-1.5%) on PACS. By contrast, our method achieves more consistent improvements, indicating the efficacy and scalability of ART. (2) ART achieves even better performance than OSDA methods (upper bound) under much more challenging settings. (3) LogitNorm [84] and Tent [82] achieve inferior performance due to the imbalance between acc_k and acc_u, showing the non-triviality of performing both unknown-aware training and test-time modification.

4.2. Generalization in Other Vision Tasks

Setup. (1) Object Detection. We introduce four datasets to form three tasks, *i.e.,* Pascal VOC [23], Clipart, Watercolor, and Comic [34] datasets. They share 6 classes, where *person* is selected as $\mathcal{C}_u^t$ and the remaining 5 classes are viewed as $\mathcal{C}_s$. The Pascal VOC2007-trainval and VOC2012-trainval datasets are combined to form the source domain, and Clipart1k, Watercolor, and Comic as used as the target domains respectively. For evaluation, we introduce four metrics: Wilderness Impact (WI) [19], Absolute Open-Set Error (AOSE) [55], mean average precision of known classes (mAP$_{\mathcal{K}}$) and average precision of unknown class (AP$_{\mathcal{U}}$). **(2) Semantic Segmentation.** GTA5 [65] and Cityscapes [18] are used as the source and target domains respectively. GTA5 is a synthetic dataset generated from Grand Theft Auto 5 game engine, while Cityscapes is collected from the street scenarios of different cities. They share 19 classes in all. According to the number of pixels per class, we use 10 classes as $\mathcal{C}_s$ and the remaining 9 classes as $\mathcal{C}_t^u$. We report the mean accuracy of all classes (mAcc), mean Intersection over Union (mIOU), acc_u and hs.

Table 5. Ablation of ART on four benchmarks. hs (%) is reported. - and + denote the removal or addition of a module respectively.

Method	PACS	Office-Home	Office-31	Digits	Avg.
ART	52.3	64.8	78.5	44.6	60.1
- $\mathcal{L}_{\text{UA}}$	39.9	62.5	69.0	20.0	47.9
- $\mathcal{L}_{\text{SCE}}$	43.4	60.0	71.5	41.3	54.1
- UGD	45.5	57.0	66.6	32.0	50.3
- TUR & $\mathcal{L}_{\text{UA}}$	44.4	61.4	65.8	7.9	44.9
- TUR & $\mathcal{L}_{\text{SCE}}$	41.0	58.9	63.0	40.3	50.8
UGD	48.1	63.7	68.2	40.9	55.2
+ TTT [76]	48.5	60.8	72.8	41.3	55.9
+ Tent [82]	37.8	45.3	64.9	33.2	45.3
+ T3A [35]	49.2	62.7	72.0	41.7	56.4
+ MEMO [95]	49.9	61.4	75.4	41.0	56.9
+ SHOT [47]	46.6	50.3	71.5	33.5	50.5
+ AaD [90]	50.2	62.5	74.7	41.8	57.3

Implementation Details. (1) Object Detection. We utilize Faster R-CNN [64] as the detection model and ResNet-50 with FPN [48] as the backbone network. To avoid the mutual influence between classification and regression heads, the original shared FC layer is replaced by two parallel FC layers. The networks are trained for 40 epochs. **(2) Semantic Segmentation.** We adopt DeepLab-v2 [16] segmentation network with ResNet-101 backbone. We use SGD optimizer with an initial learning rate of 5×10^{-4}, momentum of 0.9, and weight decay of 10^{-4}.

Results. Tab. 3 shows the detection results compared to ORE [37], OpenDet [28], and several variants of ART. With respect to mAP$_{\mathcal{K}}$ and AP$_{\mathcal{U}}$, ART outperforms the previous best method by 1.5% and 1.8% on average, revealing that ART strikes a better balance between identifications of known- and unknown-class objects. Fig. 4 provides the qualitative comparisons, where ART could precisely identify unknown samples and exhibits better bounding box regression results. For semantic segmentation, Tab. 4 reveal that even in the dense prediction task, ART is capable of significantly improving the generalization ability of deep models. The qualitative results are shown in Fig. 5, where the predictions given by ART are smoother and contain much fewer spurious areas than One Ring-S [91] and ART w/o TUR, especially on the unknown classes (*rider* and *bike*).

4.3. Discussion

Ablation study. (1) In Tab. 5, we evaluate the contribution of the different components of ART. It is evident that each of these components is reasonably designed, as the removal of any one of them leads to a commensurate reduction in accuracy. Note that when $\mathcal{L}_{\text{UA}}$ is removed, we will

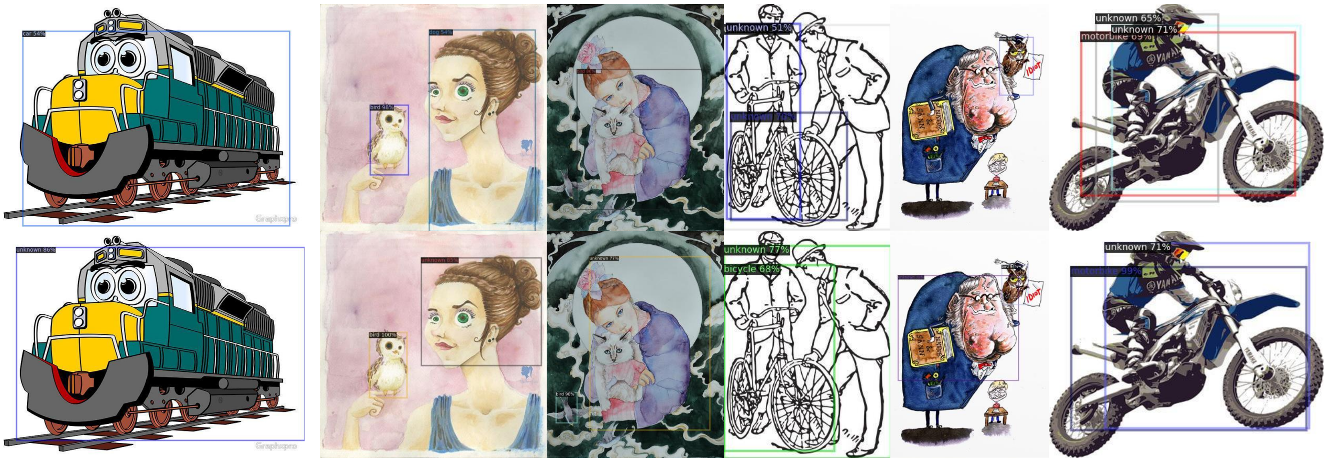
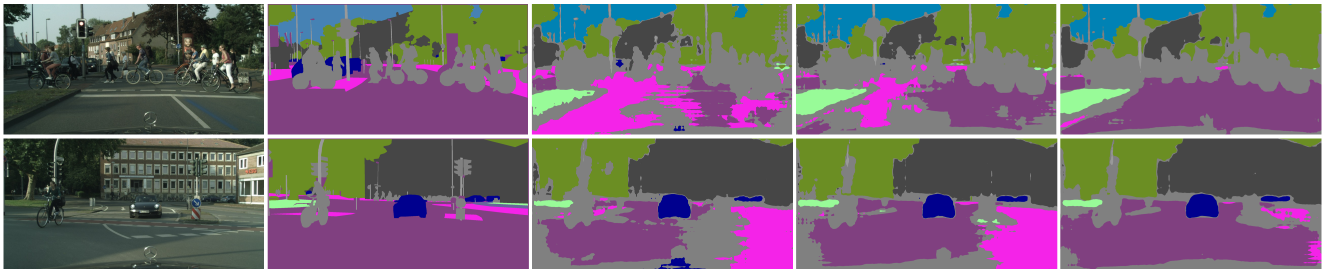
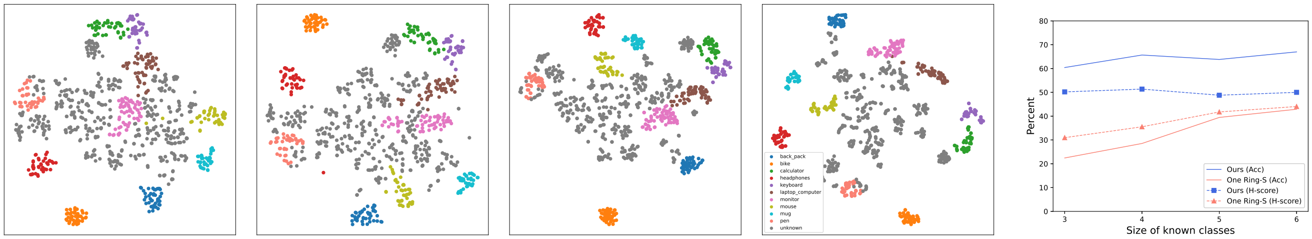


Figure 4. Qualitative comparisons between OpenDet (top) and `ART` (bottom).



(a) Target Image **(b)** Ground Truth **(c)** One Ring-S **(d)** ART w/o TUR **(e)** ART

Figure 5. Visualization of segmentation results for the task GTA5 → Cityscapes. Gray regions indicate the unknown-class pixels.



(a) ERM (b) One Ring-S (c) `ART` w/o $\mathcal{L}_{\text{SCE}}$ (d) `ART` (e) One Ring-S *vs.* `ART`

Figure 6. (a)-(d) t-SNE visualization [77] of the penultimate layer's feature on Office-31. (e) Varying the size of known classes on PACS.

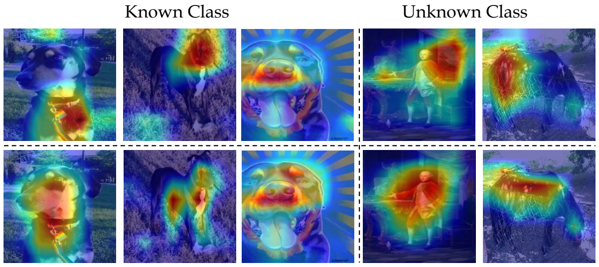


Figure 7. **Top:** w/o $\mathcal{L}_{\text{SCE}}$ *vs.* **Bottom:** w/ $\mathcal{L}_{\text{SCE}}$

Table 6. The influence of the order of test data. hs is reported.

ID	PACS	Office-Home	Office-31	Digits	Avg.
1	52.5	64.8	78.2	44.4	60.0
2	52.3	64.6	78.9	45.0	60.2
3	52.7	64.9	78.8	44.8	60.3
4	52.2	64.7	78.4	44.7	60.0

make the prediction by following the thresholding mechanism in [106]. To isolate the contribution of TUR, We additionally combine UGD with different TTA and SFDA methods. Notably, Tent and SHOT achieve inferior performance, and TTT and MEMO bring marginal improvements compared to the proposed TUR. (2) In fig. 7, we use Grad-CAM [99] to visualize the results trained w/ and w/o $\mathcal{L}_{\text{SCE}}$ on both target known- and unknown-class samples. We can observe that $\mathcal{L}_{\text{SCE}}$ makes the network focus on the entire object rather than a small or inaccurate local region, revealing the importance of mitigating the overconfidence issue in DGCS tasks.

The influence of known classes. With fixed $|\mathcal{C}_s \cup \mathcal{C}_t|$, we investigate the influence of the number of known classes. As shown in Fig. 6, ART consistently outperforms the previous best method in terms of hs especially when the size is small, indicating that ART can improve the generalization ability even with very limited known knowledge.

The influence of test order. As TUR is performed online, we study the influence of the order of test data. The results in Tab. 6 reveal that TUR is insensitive to the variations of data order, showing its robustness to the open world.

Feature visualization. We use t-SNE [77] to visualize the feature learned by ERM, One Ring-S, ART w/o $\mathcal{L}_{\text{SCE}}$, and ART, respectively. The results are displayed in Fig. 6, where different colors except for gray indicate different known classes. Points in gray represent all unknown classes. The features learned by ERM and One Ring-S cannot be reasonably separated, where the boundaries between known and unknown classes are ambiguous to some extent. By contrast, ART provides more meaningful embedding features to distinguish known and unknown samples.

5. Conclusion

We investigate the problem of DGCS, which is realistic but has been largely overlooked in the literature. Specifically, we present a simple yet surprisingly effective approach (ART) to regularize the model's decision boundary in training and adjust the source-trained classifier's prediction at test time, endowing the deep model with unknown-aware ability even without any access to real data in training. Experiments show that ART consistently improves the generalization capability of deep networks in different tasks. We hope our work will motivate future research on open-world generalization in safety-critical applications.

Acknowledgement

This work was partially supported by Hong Kong Research Grants Council under Collaborative Research Fund (Project No. HKU C7004-22G). It was also partially supported by NSFC62172348, Outstanding Young Fund of Guangdong Province with No. 2023B1515020055, and Shenzhen General Project with No. JCYJ20220530143604010.

References

[1] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *ICML*, pages 145–155, 2020. 1

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1

[3] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, pages 1563–1572, 2016. 3

[4] Silvia Bucci, Mohammad Reza Loghmani, and Tatiana Tommasi. On the effectiveness of image rotation for open set domain adaptation. In *ECCV*, 2020. 3, 6, 7

[5] Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *CVPR*, 2019. 3

[6] Chaoqi Chen, Jiongcheng Li, Xiaoguang Han, Xiaoqing Liu, and Yizhou Yu. Compound domain generalization via meta-knowledge encoding. In *CVPR*, pages 7119–7129, 2022. 2

[7] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *ICCV*, pages 2703–2712, 2021. 3

[8] Chaoqi Chen, Jiongcheng Li, Hong-Yu Zhou, Xiaoguang Han, Yue Huang, Xinghao Ding, and Yizhou Yu. Relation matters: foreground-aware graph-based relational reasoning for domain adaptive object detection. *IEEE TPAMI*, 45(3):3677–3694, 2022. 2

[9] Chaoqi Chen, Luyao Tang, Feng Liu, Gangming Zhao, Yue Huang, and Yizhou Yu. Mix and reason: Reasoning over semantic topology with data mixing for domain generalization. *NeurIPS*, 35:33302–33315, 2022. 1

[10] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *CVPR*, pages 627–636, 2019. 2

[11] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *CVPR*, pages 295–305, 2022. 3

[12] Liang Chen, Yong Zhang, Yibing Song, Anton van den Hengel, and Lingqiao Liu. Domain generalization via rationale invariance. In *ICCV*, 2023. 2

[13] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *CVPR*, pages 18710–18719, 2022. 2

[14] Liang Chen, Yong Zhang, Yibing Song, Ying Shan, and Lingqiao Liu. Improved test-time adaptation for domain generalization. In *CVPR*, pages 24172–24182, 2023. 3

[15] Liang Chen, Yong Zhang, Yibing Song, Jue Wang, and Lingqiao Liu. OST: Improving generalization of deepfake detection via one-shot test-time training. In *NeurIPS*, 2022. 2

[16] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017. 7

[17] Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE TPAMI*, 2021. 2

[18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes

dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 7

[19] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *WACV*, pages 1021–1030, 2020. 7

[20] Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *CVPR*, 2022. 3

[21] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, pages 6447–6458, 2019. 1, 2

[22] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive methods for real-world domain generalization. In *CVPR*, 2021. 2, 3

[23] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 7

[24] Bo Fu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Learning to detect open classes for universal domain adaptation. In *ECCV*, pages 567–583. Springer, 2020. 6

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

[26] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073. IEEE, 2012. 6

[27] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NeurIPS*, pages 529–536, 2005. 4

[28] Jiaming Han, Yuqiang Ren, Jian Ding, Xingjia Pan, Ke Yan, and Gui-Song Xia. Expanding low-density latent regions for open-set object detection. In *CVPR*, pages 9591–9600, 2022. 6, 7

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

[30] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021. 2

[31] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 3, 6

[32] Rui Huang and Yixuan Li. Towards scaling out-of-distribution detection for large semantic space. In *CVPR*, 2021. 3

[33] Jonathan J. Hull. A database for handwritten text recognition research. *TPAMI*, 16(5):550–554, 1994. 6

[34] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, pages 5001–5009, 2018. 7

[35] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *NeurIPS*, 34:2427–2440, 2021. 3, 7

[36] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 5

[37] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, pages 5830–5840, 2021. 6, 7

[38] Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, R Venkatesh Babu, et al. Towards inheritable models for open-set domain adaptation. In *CVPR*, pages 12376–12385, 2020. 3

[39] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 6

[40] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 6

[41] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, pages 7167–7177, 2018. 3

[42] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5542–5550, 2017. 4, 6

[43] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 1, 2

[44] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *ICCV*, pages 1446–1455, 2019. 1, 2

[45] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, pages 624–639, 2018. 2

[46] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *ICML*, pages 3915–3924, 2019. 2

[47] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, pages 6028–6039. PMLR, 2020. 3, 6, 7

[48] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 7

[49] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. *NeurIPS*, 34, 2021. 1, 2

[50] Jie Liu, Xiaoqing Guo, and Yixuan Yuan. Unknown-oriented learning for open set domain adaptation. In *ECCV*, 2022. 3

[51] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020. 3

[52] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015. 2

[53] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *ICML*, pages 7313–7324, 2021. 2

[54] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI*, 2020. 2

[55] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *ICRA*, pages 3243–3249. IEEE, 2018. 7

[56] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *CVPR*, pages 8690–8699, 2021. 2

[57] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 6

[58] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pages 427–436, 2015. 4

[59] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011. 2

[60] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *ICCV*, pages 754–763, 2017. 3

[61] Prashant Pandey, Mrigank Raman, Sumanth Varambally, and Prathosh Ap. Generalization on unseen domains via inference-time label-preserving target projections. In *CVPR*, pages 12924–12933, 2021. 3

[62] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *JMLR*, 12:2825–2830, 2011. 4

[63] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *ICML*, pages 7728–7738, 2020. 1, 2

[64] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 7

[65] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, pages 102–118. Springer, 2016. 7

[66] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *JMLR*, 19(1):1309–1342, 2018. 2

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

[68] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. *NeurIPS*, 33:16282–16292, 2020. 3, 5

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

[70] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *ICLR*, 2021. 3

[71] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *CVPR*, pages 9624–9633, 2021. 2

[72] Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-Yan Liu. Recovering latent causal factor for generalization to distributional shifts. *NeurIPS*, 34, 2021. 2

[73] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021. 3

[74] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *ECCV*, 2022. 3, 6, 7

[75] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *ICML*, 2022. 3

[76] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, pages 9229–9248. PMLR, 2020. 3, 6, 7

[77] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 8, 9

[78] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999. 6

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

[80] Riccardo Volpi and Vittorio Murino. Addressing model vulnerability to distributional shifts over image transformation sets. In *ICCV*, pages 7980–7989, 2019. 2

[81] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, pages 5334–5344, 2018. 1, 2, 6

[82] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 3, 6, 7

[83] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *TKDE*, 2022. 1

[84] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *ICML*, 2022. 6, 7

[85] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *ICCV*, 2021. 3

[86] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, pages 14383–14392, 2021. 2

[87] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. In *ICLR*, 2021. 1, 2

[88] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 3

[89] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In *NeurIPS*, 2021. 3

[90] Shiqi Yang, Yaxing Wang, Kai Wang, Shangling Jui, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. In *NeurIPS*, 2022. 3, 6, 7

[91] Shiqi Yang, Yaxing Wang, Kai Wang, Shangling Jui, and Joost van de Weijer. One ring to bring them all: Towards open-set recognition under domain shift. *arXiv preprint arXiv:2206.03600*, 2022. 2, 3, 6, 7

[92] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *CVPR*, pages 2720–2729, 2019. 3

[93] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. Towards principled disentanglement for domain generalization. In *CVPR*, pages 8024–8034, 2022. 1, 2

[94] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *CVPR*, 2018. 3

[95] Marvin Mengxin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *NeurIPS*, 2022. 3, 6, 7

[96] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *CVPR*, 2022. 2

[97] Ziyi Zhang, Weikai Chen, Hui Cheng, Zhen Li, Siyuan Li, Liang Lin, and Guanbin Li. Divide and contrast: Source-free domain adaptation via adaptive contrastive learning. In *NeurIPS*, 2022. 3

[98] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. *NeurIPS*, 33:14435–14447, 2020. 6

[99] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 8

[100] Hong-Yu Zhou, Yizhou Yu, Chengdi Wang, Shu Zhang, Yuanxu Gao, Jia Pan, Jun Shao, Guangming Lu, Kang Zhang, and Weimin Li. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nature Biomedical Engineering*, pages 1–13, 2023. 2

[101] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *TPAMI*, 2022. 1

[102] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, pages 13025–13032, 2020. 2

[103] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, pages 561–578, 2020. 2

[104] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *TIP*, 30:8008–8018, 2021. 6

[105] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2021. 1, 2

[106] Ronghang Zhu and Sheng Li. Crossmatch: Cross-classifier consistency regularization for open-set single domain generalization. In *ICLR*, 2022. 2, 3, 5, 6, 7, 8

[107] Wei Zhu, Le Lu, Jing Xiao, Mei Han, Jiebo Luo, and Adam P Harrison. Localized adversarial domain generalization. In *CVPR*, pages 7108–7118, 2022. 2