Mutual Nearest Neighbor Contrast and Hybrid Prototype Self-training for Universal Domain Adaptation

Liang Chen [1], **Qianjin Du** [2], **Yihang Lou** [2*], **Jianzhong He** [2], **Tao Bai** [2], **Minghua Deng** [1†]

[1] School of Mathematical Sciences, Peking University
[2] GoTen AI Lab, Intelligent Vision Dept, Huawei Technologies

Abstract

Universal domain adaptation (UniDA) aims to transfer knowledge learned from a labeled source domain to an unlabeled target domain under domain shift and category shift. Without prior category overlap information, it is challenging to simultaneously align the common categories between two domains and separate their respective private categories. Additionally, previous studies utilize the source classifier's prediction to obtain various known labels and one generic "unknown" label of target samples. However, over-reliance on learned classifier knowledge is inevitably biased to source data, ignoring the intrinsic structure of target domain. Therefore, in this paper, we propose a novel two-stage UniDA framework called MATHS based on the principle of **M**utual ne**A**rest neighbor con**T**rast and **H**ybrid prototype di**S**crimination. In the first stage, we design an efficient mutual nearest neighbor contrastive learning scheme to achieve feature alignment, which exploits the instance-level affinity relationship to uncover the intrinsic structure of two domains. We introduce a bimodality hypothesis for the maximum discriminative probability distribution to detect the possible target private samples, and present a data-based statistical approach to separate the common and private categories. In the second stage, to obtain more reliable label predictions, we propose an incremental pseudo-classifier for target data only, which is driven by the hybrid representative prototypes. A confidence-guided prototype contrastive loss is designed to optimize the category allocation uncertainty via a self-training mechanism. Extensive experiments on three benchmarks demonstrate that MATHS outperforms previous state-of-the-arts on most UniDA settings.

Introduction

In the past few years, deep learning has achieved impressive progress in image classification tasks. The impressive efficacy of deep learning algorithms highly relies on abundant labeled training data. However, collecting abundant labeled datasets requires massive annotation resources. A natural idea is to adapt the model well-trained on a labeled source domain to an unlabeled target domain. Unfortunately, because of the data distribution shift between different domains, the resultant model often fails to obtain acceptable generalization performance. Domain adaptation (DA) aims to exploit the learned knowledge that is transferred from source domain to target domain by eliminating domain bias (Saenko et al. 2010). Closed-set domain adaptation (CDA) assumes that the label spaces of source and target domain are consistent, in order to facilitate learning a domain-invariant representation (Ganin and Lempitsky 2015). As the restriction of label consistency is too harsh, researchers have begun to allow the existence of category shift to study partial domain adaptation (PDA) (Cao et al. 2018), open-set domain adaptation (ODA) (Saito et al. 2018) and open-partial domain adaptation (OPDA) (Fu et al. 2020). As shown in Figure 1, these three settings correspond to larger source label space, larger target label space and the partial overlapping between the two label spaces, respectively. The latter two are more challenging, because we need to align the known common categories between the two domains as well as discover the unknown private categories in the target domain.

Whether it is CDA, PDA, ODA or OPDA, we need to know the categorical relationship between source and target domain in advance, but this is not an easy task to satisfy in real cases. For more general situations, universal domain adaptation (UniDA), which imposes no prior knowledge on the target label space, was proposed (You et al. 2019). In the UniDA setting, it is extremely troublesome to align the common categories while separating the respective private categories. Excessive reliance on the supervision information of source domain would result in losing the ability to discriminate the private categories in target domain. Some previous works like UAN (You et al. 2019) and CMU (Fu et al. 2020) design sample-level transferability measure to distinguish common and private categories. They determine the category labels of target samples by means of a validated threshold. However, it is not practical to manually set a threshold to reject the target private categories. Other related works like DANCE (Saito et al. 2020) and DCC (Li et al. 2021) utilize the global clustering structure of target domain to learn a discriminative target-oriented representation. But focusing on clustering the target domain would weaken the constraint of common categories in the source domain, potentially leading to category misalignment.

To date, few feature alignment methods are specifically tailored to the UniDA setting (Saito et al. 2020). However, without domain invariant representation, adapting to the

*This work is completed in Huawei GoTen AI Lab.
†Corresponding Author.

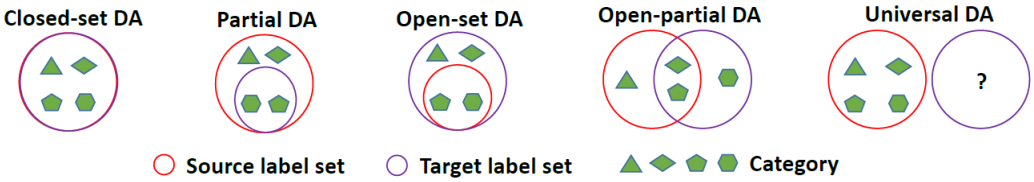


Figure 1: Various DA settings with respect to label space of source and target domains. UniDA can easily accommodate the setting where the label space of target domain is unknown.

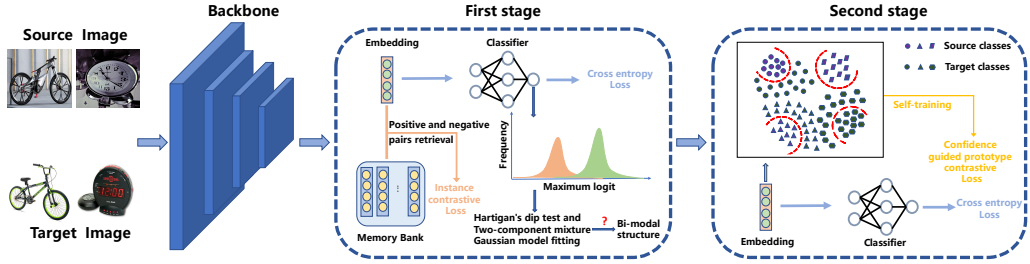


Figure 2: Schematics of our proposed two-stage UniDA framework MATHS. In the first stage, we use mutual nearest neighbors contrastive learning to achieve domain alignment. After that, we use statistical testing and fitting to detect target private samples. In the second stage, we apply a self-training strategy to obtain reliable target predictions based on hybrid prototypes.

source classifier that is biased to source data leads to highly noisy discrimination in target domain. Since target domain may contain potential private categories, global domain-level or cluster-level alignment would bring the risk of negative class knowledge transfer. Besides, most existing UniDA methods treat target private samples as a whole and ignore the class diversity of their intrinsic structure, thereby fails to learning optimal compact feature representation. More importantly, almost all UniDA algorithms cannot tell us how to distinguish between the open set and non-open set DA situations, and they use the same discrimination strategy like source classifier to deal with various situations. When "unknown" samples exist, it is suboptimal to discriminate all target samples by a closed source classifier.

To address the above issues, we propose a two-stage UniDA framework called MATHS based on the criterion of contrastive alignment of mutual nearest neighbors, and incremental pseudo-classifier discrimination guided by hybrid prototype completion. An overview of MATHS is shown in Figure 2. First, we retrieve the mutual nearest neighbor pairs both intra- and inter-domains. By eliminating the feature discrepancy of these anchor pairs, we achieve the alignment of common classes across domains and separation of private classes in each domain. This adaptation process can uncover the intrinsic structure in both two domains from the perspective of instance affinity relationship. Second, we propose to identify whether the target domain contains private categories by statistically validating whether the target discriminative probability distribution appears as a bimodal structure. We further estimate the bimodal distribution parameters and utilize them to select partial target private samples with high confidence. Third, we propose mixing the source prototypes and target private prototypes in the embedding space as an auxiliary pseudo-classifier to obtain more reliable target predictions. This process designs a self-training strategy of hybrid prototype allocation confidence.

Our contributions are summarized as follows:

- We propose a novel two-stage UniDA framework called MATHS based on mutual nearest neighbor contrast criterion and hybrid prototype discrimination principle.
- We first construct an instance-level contrastive loss to reduce the feature discrepancy between mutual nearest neighbors. Then we design a confidence-guided prototype contrastive loss to optimize the uncertainty of category allocation for target samples.
- We introduce a bimodality hypothesis for target discriminative probabilities to identify the potential target private samples, and propose a data-based statistical approach to separate the common and private categories.
- MATHS outperforms other state-of-the-art methods on challenging ODA and OPDA settings, and achieves comparable performance on CDA and PDA settings. We also conduct careful ablation studies to verify the efficacy of individual components proposed in MATHS.

Related work

Universal domain adaptation

Universal domain adaptation is a realistic but challenging DA scenario which allows both domains having their own private categories. UAN (You et al. 2019) measures the sample-level transferability to distinguish the the common and private categories. CMU (Fu et al. 2020) detects the target "unknown" samples by aggregating multiple complementary uncertainty measures. DANCE (Saito et al. 2020) designs two loss functions, neighborhood clustering and entropy separation, for category shift-agnostic adaptation. DCC (Li et al. 2021) draws the domain consensus knowledge to facilitate the target domain clustering and the private category discovery. These methods rely on source classifier to predict the target samples and barely exploit the manifold structure relationship between two domains. Conversely, MATHS achieves domain alignment by eliminating feature bias between mutual nearest neighbors, and makes reliable discrimination on target samples by hybrid prototype pseudo-classifier, instead of source classifier.

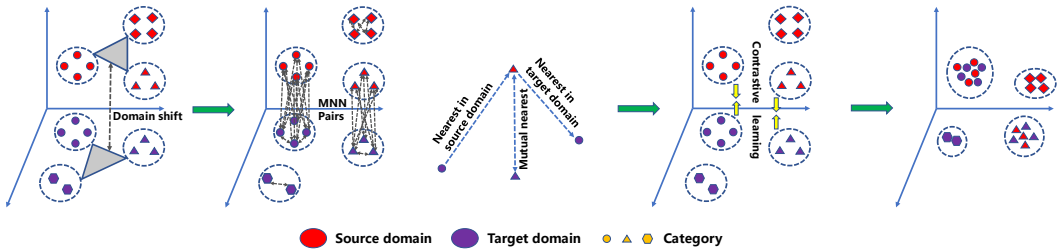


Figure 3: Schematics of our proposed universal domain alignment strategy driven by mutual nearest neighbors contrastive learning.

Out-of-distribution detection

Out-of-distribution (OOD) detection is alternatively referred to as outlier or anomaly detection, and it aims to identify whether a test example is drawn far from the train data distribution or not (Hodge and Austin 2004). Recently developed OOD detection methods mainly include those based on uncertainty measure of the classifier (DeVries and Taylor 2018), reconstruction error of the generative model (Zong et al. 2018) and self-supervised contrastive learning (Tack et al. 2020). In principle, identifying target private categories that do not belong to the source domain in UniDA is similar to OOD detection. However, applying these OOD methods to UniDA task directly may lead to erroneous identification due to domain shift. Under the premise of eliminating the domain bias, we discover that the maximum discriminative probability of the common and private classes shows a bimodal distribution, which is also the general consensus in OOD detection (Clifton, Hugueny, and Tarassenko 2011; Kamoi and Kobayashi 2020). The left peak with lower classification confidence corresponds to those target private examples. On this basis, we design a data-based statistical approach to identify partial reliable ones among them, and the remaining fuzzy ones would be detected by exploiting their affinity membership to the hybrid prototypes.

Methods

In universal domain adaptation, given a labeled source domain $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$, we need to annotate an unlabeled target domain $D_t = \{(x_i^t)\}_{i=1}^{n_t}$, which contains all, or some, known classes and possible unknown classes. Define Y_s and Y_t as the label space in source domain and target domain, respectively, and their common label space is $Y = Y_s \cap Y_t$. Note that Y may be a proper subset of Y_s, and that Y_t may contain categories that are not in Y_s. For simplicity, let $\bar{Y}_s = Y_s \setminus Y$ and $\bar{Y}_t = Y_t \setminus Y$. Our goal is to assign the label in Y to the target samples belonging to the common classes between two domains, and to assign the "unknown" label to the samples in the target private classes. Our model is a transductive transfer learning based framework, in which the labeled source domain and unlabeled target domain both participate in the network training. The overall network architecture consists of a backbone to extract embedding features of samples and a classifier to discriminate embedding features. Suppose the function for learning embedding features is $\varphi : \mathcal{X} \to \mathcal{Z}$, and the discrimination function of the classifier is $\phi : \mathcal{Z} \to \mathcal{R}^{k_s}$. The dimension k_s is the number of categories in the source domain label space Y_s.

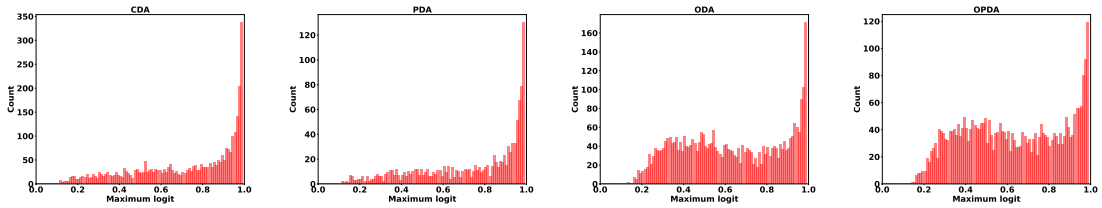
Feature alignment by mutual nearest neighbors contrastive learning paradigm

Usually, the classifier ϕ is trained on the labeled source domain using the cross-entropy loss, i.e.,

$$L_{ce} = -\frac{1}{n_s}\sum_{i=1}^{n_s}\sum_{j=1}^{k_s} y_{i,j}^s \log(\phi(\varphi(x_i^s))_j). \tag{1}$$

Because of the domain shift, the classifier trained on source domain cannot generalize well to target domain. Many methods rely on global distribution calibration to learn domain invariant feature representation, such as those based on adversarial learning (Ganin and Lempitsky 2015) and maximum mean discrepancy (Yan et al. 2017). However, they do not consider class specific structure properties, and this may result in noisy prediction near the classification boundary. Since the target domain may contain categories not in the source domain, class-level domain alignment methods, like those based on pseudo-labeling (Liang, Hu, and Feng 2020) and clustering (Tang, Chen, and Jia 2020), may encounter categories misalignment issue, leading to negative transfer. Therefore, inspired by manifold learning (McInnes, Healy, and Melville 2018), we exploit the much finer domain structure knowledge from the perspective of nearest neighbor graph and pairwise affinity constraint. For different domains, the geometrically nearest neighbors can be considered as the most similar anchor pairs in the same category. For each domain itself, especially unlabeled target domain, geometrically close samples are more likely to belong to the same category. To force alignment of both two domains on the common categories and compact on the respective private categories, we propose a contrastive learning framework based on mutual nearest neighbors (see Figure 3).

Specifically, for any sample i in target domain, we search its k nearest neighbors G_i^s in the source domain; similarly, for the sample j in source domain, we can also obtain its k nearest neighbors G_j^t in the target domain. If an instance pair from different domains is contained in each other's nearest neighbor set, namely $i \in G_j^t, j \in G_i^s$, they are considered to be mutual nearest neighbors. On this basis, we construct a pairwise affinity relationship between two domains. We build this relationship as an adjacency matrix $A^{st} \in R^{n_s \times n_t}$. Then $A_{ij}^{st} = 1$ if and only if i and j is the mutual nearest neighbor pair, or positive pair; otherwise $A_{ij}^{st} = 0$, or negative pair. To explore the target domain's intrinsic structure, we also search the mutual nearest neighbor pairs in it and construct its adjacency matrix $A^{tt} \in R^{n_t \times n_t}$. Since the source domain has ground-truth labels, its affinity matrix $A^{ss} \in R^{n_s \times n_s}$ can be obtained according to whether the sample pair belongs to the same class.



(a) CDA (b) PDA (c) ODA (d) OPDA

Figure 4: The frequency histogram of $\{q_i\}_{i=1}^{n_t}$ on "Art (source) to Real (target)" in OfficeHome (Venkateswara et al. 2017). (a)(b)No significant bimodal structure; (c)(d)Significant bimodal structure.

In the training phase, we aim to pull these nearest neighbors closer to each other and push away those geometrically dissimilar samples to prevent category structure collapse. To accomplish it, we design an instance contrastive loss based on the affinity matrices. Considering the limited information contained in the minibatch, we employ an effective memory bank $\mathcal{B}$ (Zhuang, Zhai, and Yamins 2021) to store diverse global information of two domains. The bank $\mathcal{B}$ contains both updated source and target features from the current minibatch and the older features absent in the minibatch, without utilizing the exponential moving average of features in previous epochs. For each training sample i, its positive and negative sample sets in $\mathcal{B}$ are $\mathcal{P}_i$ and $\mathcal{N}_i$, respectively, which can be inferred by affinity matrices A^{ss}, A^{st} and A^{tt}. Inspired by InfoNCE (Hadsell, Chopra, and LeCun 2006), our instance contrastive loss is

$$L_{Ins-Con} = -\sum_{i=1}^{n_s+n_t}\sum_{j\in\mathcal{P}_i}\log\psi_{i,j}, \tag{2}$$

$$\psi_{i,j} = \frac{\exp(z_i z_j/\tau)}{\sum_{k\in\mathcal{P}_i}\exp(z_i z_k/\tau)+\sum_{l\in\mathcal{N}_i}\exp(z_i z_l/\tau)},$$

where τ is a temperature parameter. By minimizing $L_{ce} + \lambda L_{Ins-Con}$ (λ is a weight parameter), we strive to spread out the known categories in the source domain, align the common categories between two domains, and compact the clusters in the target domain.

Separating common and private categories with bimodality hypothesis

After contrastive feature alignment on both two domains, we transform the recognition of open set DA into the recognition of ODD detection. The extreme value theory (Clifton, Hugueny, and Tarassenko 2011) points out that OOD detection in multivariate data is equivalent to performing novelty detection in the probability space of the model of normality. Given the discriminative probability vectors $\{p_i\}_{i=1}^{n_t}$ of target samples, we would typically take the category index corresponding to their maximum logit $q_i = \max(p_i)$ as the predicted label. Since the embedding features are provided by the same engine trained on "known" categories, the logits $\{q_i\}_{i=1}^{n_t^c}$ of the common classes should be higher than those of private classes (RoyChowdhury et al. 2020). Equivalently, when the target domain contains private classes, the distribution of $\{q_i\}_{i=1}^{n_t}$ would show a bimodal structure (see Figure 4). To identity the target private samples, an alternative approach is to set a threshold δ for $\{q_i\}_{i=1}^{n_t}$. If $q_i < \delta$, sample i can be regarded as a target private sample.

Therefore, we first apply the Hartigan's dip test (Hartigan and Hartigan 1985) to validate the bimodality of $\{q_i\}_{i=1}^{n_t}$. The dip test measures multimodality in a sample by the maximum difference, over all sample points, between the empirical distribution function, and the unimodal distribution function that minimizes that maximum difference. When the test p-value is small (< 0.05), we can determine that the target domain contains private categories. Then we use a two-component Gaussian mixture model $\pi N(\theta_1, \sigma_1^2) + (1-\pi)N(\theta_2, \sigma_2^2), \theta_1 < \theta_2$ to fit the distribution of $\{q_i\}_{i=1}^{n_t}$. According to the three-sigma rule, we take the "unknown" threshold as $\hat{\delta} = \hat{\theta}_1 + \hat{\sigma}_1$ default which can contain about 80% of samples in the left peak. We classify the samples with $q_i < \hat{\delta}$ into target private categories. These samples will be given the "unknown" label and denoted as D_t^p.

Reliable prediction via hybrid prototype completion and self-training

In the previous section, we select partial target samples as representative instances of private classes. We further assume that they cover all private classes in the target domain. Then we abandon the noisy prediction of source classifier on remaining fuzzy target samples. Different from other methods, a pseudo-classifier discrimination method is proposed in the embedding space, which combines the source prototypes and target private prototypes to make a reliable prediction. Our motivation for using prototypes as the pseudo-classifier is that the prototypes can denoise the false target labels near the classifier boundary by weakening the contribution of outliers. Thus, we first perform k-means clustering with cluster number k_p on embedding features of the samples in D_t^p. Assuming that the source class centroids and target cluster centroids are $\{\mu_j^s\}_{j=1}^{k_s}$ and $\{\nu_j^t\}_{j=1}^{k_p}$, respectively, then we aim to move all target samples to a geometrically close centroid through self-training. Specifically, the samples in D_t^p are moved to the vicinity of corresponding target private cluster centroid ν, and the samples in $D_t \setminus D_t^p$ are moved to the vicinity of μ or ν through the principle of minimizing allocation uncertainty. We propose a confidence-guided prototype contrastive loss to unify these objectives. Assuming that all embedding features and class centroids are l_2 normalized, we have

$$L_{Pro-Con} = -\sum_{i=1}^{n_t}\sum_{j=1}^{k_s+k_p} w_{i,j}\log s_{i,j}, \tag{3}$$

$$s_{i,j} = \frac{\exp(z_i\varepsilon_j/\tau)}{\sum_{k=1}^{k_s}\exp(z_i\mu_k^s/\tau)+\sum_{l=1}^{k_p}\exp(z_i\nu_l^t/\tau)},$$

$$\varepsilon_j = \mu_j^s, 1 \le j \le k_s,$$

$$\varepsilon_j = \nu_{j-k_s}^t, k_s+1 \le j \le k_s+k_p,$$

Table 1: Results comparison on closed-set domain adaptation (CDA).

Universal comparison																					
	Office (31/0/0)							OfficeHome (65/0/0)													VisDA (12/0/0)
Methods	A2W	D2W	W2D	A2D	D2A	W2A	Avg	A2C	A2P	A2R	C2A	C2P	C2R	P2A	P2C	P2R	R2A	R2C	R2P	Avg	S2R
SO	74.1	95.3	99.0	80.1	54.0	56.3	76.5	37.0	62.2	70.7	46.6	55.1	60.3	46.1	32.0	68.7	61.8	39.2	75.4	54.6	46.3
DANN	86.7	97.2	99.8	86.1	72.5	72.8	85.9	46.8	68.4	76.6	54.7	63.9	69.7	57.1	44.7	75.7	64.9	51.3	78.7	62.7	69.1
ETN	87.9	**99.2**	**100**	88.4	68.7	66.8	85.2	46.7	69.5	74.8	62.1	66.9	71.9	56.7	44.1	77.0	70.6	50.4	77.9	64.0	64.1
STA	77.1	90.7	98.1	75.5	51.4	48.9	73.6	30.4	46.8	55.9	33.6	46.2	51.1	35.0	28.3	58.2	51.3	33.1	66.5	44.7	48.1
UAN	86.5	97.0	**100**	84.5	69.6	68.7	84.4	45.0	63.6	71.2	51.4	58.2	63.2	52.6	40.9	71.0	63.3	48.2	75.4	58.7	66.4
CMU	79.6	98.1	97.6	78.3	62.3	63.4	79.9	42.8	65.6	74.3	58.1	63.1	67.4	54.2	41.2	73.8	66.9	48.0	78.7	61.2	56.9
DANCE	88.6	97.5	**100**	89.4	69.5	68.2	85.5	54.3	75.9	**78.4**	64.8	72.1	73.4	63.2	53.0	**79.4**	73.0	58.2	82.9	69.1	70.2
DCC	89.1	96.8	**100**	87.2	**74.4**	**76.8**	**87.4**	35.4	61.4	75.2	45.7	59.1	62.7	43.9	30.9	70.2	57.8	41.0	77.9	55.1	69.3
MATHS	**90.4**	97.8	**100**	**90.7**	71.6	70.5	86.8	**54.7**	**76.3**	78.0	**65.4**	**73.5**	**74.6**	**64.8**	**55.7**	78.8	**73.8**	**59.7**	**83.4**	**69.9**	**72.9**
Methods tailored for Closed-set Domain Adaptation																					
CDAN	93.1	98.2	100	89.8	70.1	68.0	86.6	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8	70.0
MDD	94.5	98.4	100	93.5	74.6	72.2	88.9	54.9	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	60.2	82.3	68.1	74.6

Table 2: Results comparison on partial domain adaptation (PDA).

Universal comparison																					
	Office-Caltech (10/21/0)						OfficeHome (25/40/0)														VisDA (6/6/0)
Methods	A2C	W2C	D2C	D2A	W2A	Avg	A2C	A2P	A2R	C2A	C2P	C2R	P2A	P2C	P2R	R2A	R2C	R2P	Avg	S2R
SO	75.4	70.7	68.5	80.4	84.6	75.9	37.1	64.5	77.1	52.0	51.3	62.4	52.0	31.3	71.6	66.6	42.6	75.1	57.0	46.3
DANN	41.9	42.7	43.4	41.5	41.5	42.2	35.5	48.2	51.6	35.2	35.4	41.4	34.8	31.7	46.2	47.5	34.7	49.0	40.9	38.7
ETN	88.9	**92.3**	**92.9**	**95.4**	94.3	**92.8**	52.1	74.5	83.1	69.8	65.2	76.5	69.1	50.6	82.5	76.3	53.8	79.1	69.4	59.8
STA	75.7	72.4	62.8	70.5	67.7	69.8	35.0	55.2	59.7	37.5	48.4	53.5	36.0	32.2	59.9	54.3	38.5	64.6	47.9	48.2
UAN	47.1	49.7	50.6	55.5	61.6	52.9	24.5	35.0	41.5	34.7	32.3	32.7	32.7	21.1	43.0	39.7	26.6	46.0	34.2	39.7
CMU	56.3	60.8	63.7	69.2	66.8	63.4	50.9	74.2	78.4	62.2	64.1	72.5	63.5	47.9	78.3	72.4	54.7	78.9	66.5	65.5
DANCE	88.8	79.2	79.4	83.7	92.6	84.8	53.6	73.2	84.9	70.8	67.3	82.6	70.0	50.9	**84.8**	77.0	55.9	81.8	71.1	73.7
DCC	85.7	83.4	82.9	**95.4**	**95.5**	88.6	54.2	47.5	57.5	**83.8**	**71.6**	86.2	63.7	**65.0**	75.2	**85.5**	**78.2**	82.6	70.9	72.4
MATHS	**89.1**	86.2	85.7	85.4	92.8	87.9	**56.3**	**74.8**	**85.6**	71.2	69.4	83.5	**70.6**	52.7	83.6	76.5	57.3	**82.9**	**72.0**	**74.8**
Methods tailored for Partial Domain Adaptation																				
IFAN	NA	NA	NA	NA	NA	NA	58.9	76.3	81.4	70.4	73.0	77.8	72.4	55.3	80.4	75.8	60.4	79.9	71.8	67.7
BA^{3}US	91.8	93.5	93.9	94.8	95.0	93.8	60.6	83.2	88.4	71.8	72.8	83.4	75.5	61.6	86.5	79.3	62.8	86.1	75.9	54.9

where w is an expanded soft label vector. For the sample b in D_t^p, suppose it belongs to the r-th item of $\{\nu_j^t\}_{j=1}^{k_p}$, then $w_{b,k_s+r} = 1$ and $w_{b,k} = 0$ ($k \neq k_s + r, 1 \leq k \leq k_s + k_p$). For the sample c in $D_t \setminus D_t^p$, we take $w_{c,j} = \frac{s_{c,j}^2 / \sum_{l=1}^{n_t} s_{l,j}}{\sum_{k=1}^{k_s+k_p} s_{c,k}^2 / \sum_{l=1}^{n_t} s_{l,k}}$ ($1 \leq j \leq k_s + k_p$). The principle of designing $w_{i,j}$ is to make its distribution sharper and more concentrated than that of contrastive score $s_{i,j}$. Minimizing $L_{Pro-Con}$ is a self-training process, which can improve cluster purity and put more emphasis on samples assigned with high confidence to strengthen predictions. As for the update of class centroids $\{\mu_j^s\}_{j=1}^{k_s} \cup \{\nu_j^t\}_{j=1}^{k_p}$, we set them as variables and automatically update them using the back-propagation gradient optimized by the network. Their initialization can be obtained after the first stage of training. For known categories in the source domain, we can directly use their labels and embedding information to calculate the corresponding initialized class centroids $\{\mu_j^{s,initial}\}_{j=1}^{k_s}$. For target private categories, k-means algorithm would output the initialized cluster centroids $\{\nu_j^{t,initial}\}_{j=1}^{k_p}$. When the Hartigan's dip test presents no significant bimodal distribution, we will only use the source prototypes $\{\mu_j^s\}_{j=1}^{k_s}$ to perform self-training and inference. The final predicted label will be determined by the index corresponding to the maximum component of contrastive score vector s_i.

Experiments

Setup

Datasets. We conduct experiments on three benchmark datasets: Office (Saenko et al. 2010), OfficeHome (Venkateswara et al. 2017) and VisDA (Peng et al. 2017). Office contains three domains (Amazon (A), DSLR (D), Webcam (W)) and consists of 4652 images from 31 classes. OfficeHome is a more challenging dataset with 15500 images from 65 classes, and consists of four domains (Artistic images (A), Clip-Art images (C), Product images (P), and Real world images (R)). VisDA has 12 classes from two domains: the source domain contains 150000 synthetic images (S) and the target domain consists of 50000 real world images (R). We also use the Caltech dataset (Griffin, Holub, and Perona 2007) in PDA setting similar to that in DANCE. Let $|Y|$, $|\hat{Y}_s|$ and $|\hat{Y}_t|$ denote the number of common classes, source private classes and target private classes, respectively. We show the class split ($|Y|/|\hat{Y}_s|/|\hat{Y}_t|$) of each experimental setting in corresponding result table. The split details can be seen in the supplemental material.

Evaluation protocols. In CDA and PDA settings, we calculate the classification accuracy on the whole target samples. In ODA and OPDA settings, all samples in target private classes are regarded as one "unknown" class, and we report the average accuracy over the $|Y| + 1$ classes. We also use the H-score (Fu et al. 2020), the harmonic mean of the accuracy on common classes and accuracy on the "unknown" class, to evaluate each method. For all experiments, we assume no prior information about category shift in advance and report the averaged results of three runs.

Implementation details. We conduct all the experiments on 8 Tesla V100 GPUs with PyTorch (Paszke et al. 2017) implementation. The network backbone is ResNet50 (He et al. 2016) pretrained on ImageNet (Deng et al. 2009), and the classifier is made of one fully connected layer. We use the Nesterov momentum SGD with momentum 0.9 and weight decay 5e-4 to optimize our model. Similar to domain specific batch normalization, we split source and target samples into different mini-batches with size 36 and forward them separately. For the sake of consistency and fairness, the nearest neighbors number k is set to 30 and the k-means cluster number k_p is set to 10 as default. Following previous work (Saito et al. 2020), the temperature parameter τ is

Table 3: Results comparison on open-set domain adaptation (ODA).

Universal comparison																					
	Office (10/0/11)							OfficeHome (15/0/50)													VisDA (6/0/6)
Methods	A2W	D2W	W2D	A2D	D2A	W2A	Avg	A2C	A2P	A2R	C2A	C2P	C2R	P2A	P2C	P2R	R2A	R2C	R2P	Avg	S2R
SO	83.8	95.3	95.3	89.6	85.5	84.9	89.1	55.1	79.8	87.2	61.8	66.2	76.6	63.9	48.5	82.4	75.5	53.7	84.2	69.6	43.3
DANN	87.6	90.5	91.2	88.7	87.4	87.0	88.7	62.1	78.0	86.4	75.5	72.0	79.3	68.8	52.5	82.7	76.1	58.0	82.7	72.8	48.2
ETN	86.7	90.0	90.1	89.1	86.7	86.6	88.2	58.2	79.9	85.5	67.7	70.9	79.6	66.2	54.8	81.2	76.8	60.7	81.7	71.9	51.7
STA	91.7	94.4	94.8	90.9	87.3	80.6	89.9	56.6	74.7	86.5	65.7	69.7	77.3	63.4	47.8	81.0	73.6	57.1	78.8	69.3	57.1
UAN	88.0	95.8	94.8	88.1	89.9	89.4	91.0	63.3	82.4	86.3	75.3	76.2	82.0	69.4	58.2	83.4	76.1	60.5	81.9	74.6	50.0
CMU	90.4	96.7	95.9	91.6	89.3	88.2	92.0	57.6	78.7	85.9	66.1	71.8	80.6	65.8	53.9	84.1	75.6	60.1	84.0	72.0	59.3
DANCE	93.6	97.0	97.1	95.7	91.0	90.3	94.1	64.1	84.1	**88.3**	76.7	80.7	84.9	77.6	62.7	**85.4**	80.8	65.1	87.1	78.1	65.3
DCC	92.9	96.5	96.7	94.3	**91.2**	90.7	93.7	66.8	82.9	86.1	64.5	71.2	78.0	65.7	61.5	83.5	73.3	65.8	83.3	73.6	68.8
MATHS	**94.5**	**97.8**	**98.6**	**96.9**	90.7	**91.6**	**95.0**	**73.6**	**85.0**	87.7	**78.5**	**81.3**	**85.6**	**79.3**	**74.8**	85.1	**82.1**	**75.9**	**88.4**	**81.4**	**70.6**

Table 4: Results comparison on open-partial domain adaptation (OPDA).

Universal comparison																					
	Office (10/10/11)							OfficeHome (10/5/50)													VisDA (6/3/3)
Methods	A2W	D2W	W2D	A2D	D2A	W2A	Avg	A2C	A2P	A2R	C2A	C2P	C2R	P2A	P2C	P2R	R2A	R2C	R2P	Avg	S2R
SO	75.7	95.4	95.2	83.4	84.1	84.8	86.4	50.4	79.4	90.8	64.9	66.1	79.9	71.6	48.5	87.6	77.8	52.1	82.8	71.0	38.8
DANN	87.6	90.5	91.2	88.7	87.4	87.0	88.7	59.9	80.6	89.8	77.5	73.3	86.4	78.5	61.5	88.5	80.3	62.1	82.4	76.7	50.6
ETN	89.1	90.6	90.9	86.3	86.4	86.5	88.3	58.2	78.5	89.1	77.2	69.3	87.5	77.0	56.0	88.2	77.5	58.4	83.0	75.0	66.6
STA	85.2	96.3	95.1	88.1	87.9	86.0	89.8	54.8	76.6	91.2	71.5	71.8	82.0	70.7	50.1	88.2	74.1	60.0	80.5	72.6	47.4
UAN	76.2	82.0	80.4	80.0	**93.8**	92.2	84.1	60.8	79.1	87.8	72.4	73.5	83.2	78.6	56.4	87.4	79.9	61.1	79.8	75.0	47.3
CMU	86.9	95.7	98.0	89.1	88.4	88.6	91.1	63.5	83.8	88.9	77.7	79.4	86.9	78.6	59.3	88.3	84.1	64.6	81.4	78.0	61.4
DANCE	**92.8**	97.8	97.7	91.6	92.2	91.4	93.9	64.1	84.3	91.2	84.3	78.3	**89.4**	83.4	63.6	91.4	83.3	63.9	86.9	80.4	69.2
DCC	91.7	94.5	96.2	**93.7**	90.4	92.0	93.1	74.7	82.7	**92.1**	70.7	79.5	87.1	81.5	66.8	**92.1**	82.4	62.1	87.3	79.9	64.2
MATHS	92.6	**98.4**	**98.6**	93.5	93.0	**92.3**	**94.7**	**76.9**	**85.2**	91.8	**85.1**	**84.6**	89.2	**84.8**	**77.9**	91.0	**84.5**	**76.7**	**87.9**	**84.6**	**72.8**
Methods tailored for Open-partial Domain Adaptation																					
USFDA	85.6	95.2	97.8	88.5	87.5	86.6	90.2	63.4	83.3	89.4	71.0	72.3	86.1	78.5	60.2	87.4	81.6	63.2	88.2	77.0	63.9

set to 0.05. We set the weight λ as 0.5 to balance each loss contribution, which is a common choice in the community.

Baseline. We focus our comparison with previous state-of-the-art methods in four possible scenarios of UniDA, i.e. CDA (DANN (Ganin and Lempitsky 2015)), PDA (ETN (Cao et al. 2019)), ODA (STA (Liu et al. 2019)) and OPDA (UAN (You et al. 2019), CMU (Fu et al. 2020), DANCE (Saito et al. 2020), DCC (Li et al. 2021)). We also report the performance of other related methods tailed to each domain adaptation setting, such as CDAN (Long et al. 2018), MDD (Zhang et al. 2019), IFAN (Xu et al. 2019), BA^{3}US (Liang et al. 2020) and USFDA (Kundu, Venkat, and Babu 2020). The summary of these comparison methods can be seen in supplemental material. Since we cannot know the category shift in advance, we perform Hartigan's dip test and mixture Gaussian model fitting to discover the potential "unknown" samples in all experiments. The results of test p-value for each setting can be found in supplemental material.

Results

CDA setting. From the results in Table 1, MATHS shows consistently better performance than other baseline methods on the OfficeHome and VisDA datasets. When compared to those methods customized for CDA setting, MATHS also achieves the best performance in OfficeHome and competitive performance in Office and VisDA.

PDA setting. The results in Table 2 tell us that MATHS outperforms all baseline methods and even those methods specialized in this setting on the large-scale VisDA dataset. Although ETN shows the best performance on the Office dataset, it does not give acceptable results in either ODA or OPDA settings where "unknown" samples exist.

ODA setting. In ODA and OPDA settings, we activate the "unknown" label mechanism since the test p-values are lower than 0.05 (see supplemental material). From the results of three datasets in Table 3, MATHS consistently performs better than all baseline methods including STA tailed for ODA setting, validating the efficacy of our private sample detection approach. Besides, the H-score comparison results in supplemental material also support our claim.

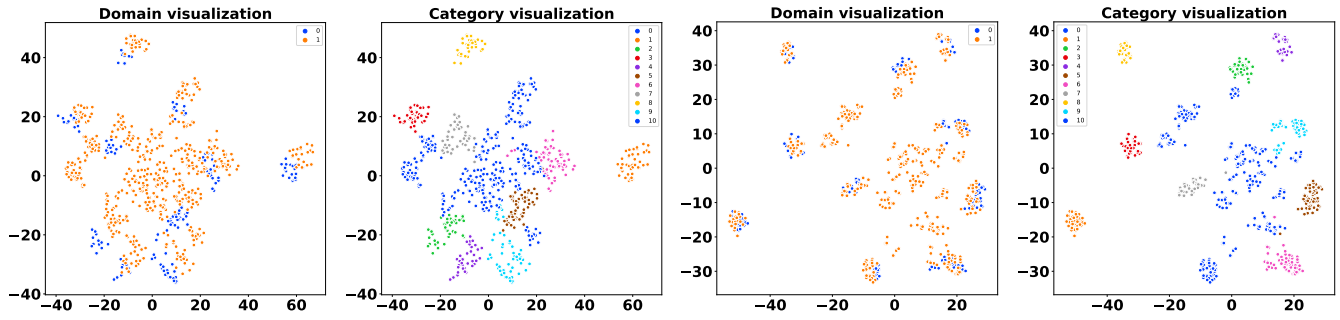
OPDA setting. In this most challenging case, the average accuracy results in Table 4 and the H-score results in supplemental material show that MATHS also consistently outperforms DANCE and achieves state-of-the-art results on all datasets. This can be attributed to its better domain alignment via contrastive learning on the mutual nearest neighbors. Instead of treating "unknown" samples as one generic class, we cluster them as additional prototypes that possess the same contribution as that of source prototypes.

Clustering the "unknown" examples. We evaluate the quality of learned embedding features by clustering the samples from both common and private classes. In the previous evaluation of ODA and OPDA settings, we classify the examples from private classes into an "unknown" group. Here we demonstrate the ability to classify them into respective private classes. Specifically, we use one labeled example per class to train a new linear classifier on the fixed embedding features. Then we calculate the classification accuracy for both common and private classes. In this experiment, we use the OfficeHome dataset in ODA setting that contains 15 common categories and 50 target private categories. Table 5 shows that MATHS improves the accuracy for both common and private classes compared to DANCE, illustrating that the embedding features learned by MATHS are more discriminative and better for separating the "unknown" group.

Table 5: Linear classification accuracy given one labeled target sample per class on OfficeHome dataset in open-set setting (Known Accuracy/ Novel Accuracy).

	R2A	R2C	R2P	P2A	P2C	P2R
Methods	known/novel	known/novel	known/novel	known/novel	known/novel	known/novel
ImageNet	37.5/31.0	35.3/36.4	64.8/56.9	36.9/31.0	36.3/36.0	66.3/45.5
SO	42.4/30.7	43.4/33.8	69.9/53.8	38.6/30.1	37.0/32.2	65.1/39.1
DANN	41.3/30.2	42.4/33.4	62.8/50.7	41.6/28.9	40.1/31.6	67.2/38.8
DANCE	49.1/33.8	48.7/36.5	74.9/57.9	46.4/35.2	43.0/38.1	74.1/45.2
MATHS	**52.8/36.1**	**54.3/39.9**	**76.0/59.6**	**49.7/37.0**	**47.6/41.8**	**74.5/46.4**

Feature visualization. We use t-SNE (Van der Maaten and Hinton 2008) to visualize the learned source and target features with corresponding domain label and category label before and after adaptation. In this analysis, we conduct ex-



(a) Before model adaptation (b) After model adaptation

Figure 5: Feature visualization on "D2W" in Office under the ODA setting. For domain, blue represents the source domain and orange refers to the target domain. For category, blue plots are "unknown" samples, others are "known" samples.

periment on "D2W" in Office under the ODA setting. From Figure 5, before adaptation, the common categories do not mix well and most target private samples are attached near the common samples. After adaptation, we achieve the common categories mixing well and separate most target private samples from the common samples. This qualitative evidence demonstrates that our cross-domain contrastive learning strategy is very effective for feature alignment.

Ablation study

Effect of mutual nearest neighbors contrast. To verify the necessity of the proposed mutual nearest neighbors (mnn) contrastive learning for feature alignment, we use VisDA dataset to conduct control experiments in four settings. From the results in Table 6, removing contrastive learning on mnn pairs would only severely degrade performance. Besides, we discover that removing it would also make the bimodal structure of maximum logits not so significant (see supplemental material). Furthermore, we replace our mnn contrastive learning paradigm with MoCo (He et al. 2020) in the first stage and test its result on the VisDA dataset. The average accuracy of MoCo in CDA, PDA, ODA, and OPDA settings is 68.5, 67.4, 63.1, and 67.2, respectively, about 6% lower than MATHS. This is reasonable since MoCo uses contrastive loss at the single instance level and treats those informative intra- and inter-domain mnn pairs as noisy negative samples. MATHS pulls these anchor pairs closer to each other by contrasting multiple positives with multiple negatives, so as to guarantee class-specific alignment. All this evidence demonstrates that using only the domain specific batch normalization and ordinary contrastive learning would be too weak to eliminate domain shift, while our proposed criterion is simple, yet effective.

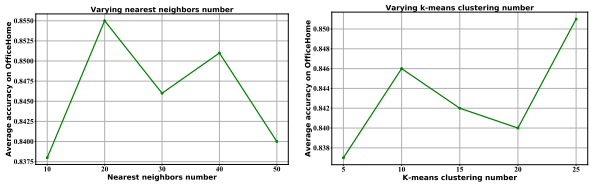
Effect of hybrid prototype self-training. To evaluate the contribution of hybrid prototype self-training to reliable prediction, we abandon it in the second stage and also use the VisDA dataset in four settings to conduct analysis. Without the self-training on hybrid prototypes, the results in Table 6 tell us that the accuracy would decline, which illustrates that self-training can make fuzzy discrimination boundary distinct and further compact feature representation. To testify the effectiveness of hybrid prototype prediction, we utilize the source classifier instead of it. Table 6 shows that in the ODA and OPDA settings, the use of source classifier to discriminate target samples is not as good as hybrid prototype prediction, validating our original claim.

To show that applying OOD methods to a UniDA setting directly may lead to erroneous identification, we only use the source classifier with our learned "unknown" threshold to test on VisDA dataset. Its average accuracy in ODA and OPDA settings is 47.3 and 45.1, respectively, much lower than MATHS, which demonstrates that the domain bias seriously harms the accuracy of the OOD methods.

Sensitivity to hyper-parameter. To show the sensitivity of MATHS to the nearest neighbors number k and k-means cluster number k_p, we conduct experiments on OfficeHome under the OPDA setting, and present the average accuracy over twelve transfer tasks on four domains, as shown in Figure 6(a) and Figure 6(b). Within a wide range of k and k_p, the average accuracy varies slightly, demonstrating that MATHS is robust to the choices of k and k_p.

Table 6: Ablation study on VisDA. mnnc, hps and hpp refer to mutual nearest neighbors contrast, hybrid prototype self-training and hybrid prototype prediction, respectively.

	CDA	PDA	ODA	OPDA
MATHS w/o mnnc	65.2	64.7	59.3	64.1
MATHS w/o hps	71.5	73.9	68.5	71.2
MATHS w/o hpp	71.8	74.0	66.4	69.7
MATHS (full)	72.9	74.8	70.6	72.8



(a) Neighbors number k (b) Cluster number k_p

Figure 6: (a) Varying nearest neighbors number k. (b) Varying k-means cluster number k_p.

Conclusion

In this paper, we propose a novel two-stage framework called MATHS for universal domain adaptation. It performs feature alignment via mutual nearest neighbors contrast and exploits domain discrimination knowledge by hybrid prototype self-training. We also introduce a data-based statistical method to detect target private categories. A thorough evaluation shows that MATHS outperforms other state-of-the-art UniDA methods on most DA settings. Beyond the UniDA task, MATHS shows promise for unsupervised transfer learning and data integration tasks, which are outside the scope of this paper, but need to be explored.

References

Cao, Z.; Long, M.; Wang, J.; and Jordan, M. 2018. Partial transfer learning with selective adversarial networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2724–2732.

Cao, Z.; You, K.; Long, M.; Wang, J.; and Yang, Q. 2019. Learning to transfer examples for partial domain adaptation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2985–2994.

Clifton, D. A.; Hugueny, S.; and Tarassenko, L. 2011. Novelty detection with multivariate extreme value statistics. *Journal of signal processing systems*, 371–389.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Li, F.-F. 2009. Imagenet: A large-scale hierarchical image database. *In 2009 IEEE conference on computer vision and pattern recognition*, 248–255.

DeVries, T.; and Taylor, G. 2018. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*.

Fu, B.; Cao, Z.; Long, M.; and Wang, J. 2020. Learning to Detect Open Classes for Universal Domain Adaptation. *In European Conference on Computer Vision*, 567–583.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. *In International conference on machine learning*, 1180–1189.

Griffin, G.; Holub, A.; and Perona, P. 2007. Caltech-256 object category dataset. *California Institute of Technology*.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1735–1742.

Hartigan, J. A.; and Hartigan, P. M. 1985. The dip test of unimodality. *Annals of statistics*, 13(1): 70–84.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hodge, V.; and Austin, J. 2004. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2): 85–126.

Kamoi, R.; and Kobayashi, K. 2020. Out-of-Distribution Detection with Likelihoods Assigned by Deep Generative Models Using Multimodal Prior Distributions. *In SafeAI@ AAAI*, 113–116.

Kundu, J. N.; Venkat, N.; and Babu, R. V. 2020. Universal source-free domain adaptation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4544–4553.

Li, G.; Kang, G.; Zhu, Y.; Wei, Y.; and Yang, Y. 2021. Domain Consensus Clustering for Universal Domain Adaptation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Liang, J.; Hu, D.; and Feng, J. 2020. Domain Adaptation with Auxiliary Target Domain-Oriented Classifier. *arXiv preprint arXiv:2007.04171*.

Liang, J.; Wang, Y.; Hu, D.; He, R.; and Feng, J. 2020. A balanced and uncertainty-aware approach for partial domain adaptation. *In Computer Vision–ECCV 2020: 16th European Conference*, 123–140.

Liu, H.; Cao, Z.; Long, M.; Wang, J.; and Yang, Q. 2019. Separate to adapt: Open set domain adaptation via progressive separation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2927–2936.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. *In Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 1647–1657.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

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

RoyChowdhury, A.; Yu, X.; Sohn, K.; Erik, L.-M.; and Manmohan, C. 2020. Improving Face Recognition by Clustering Unlabeled Faces in the Wild. *In European Conference on Computer Vision*, 119–136.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. *In European conference on computer vision*, 213–226.

Saito, K.; Kim, D.; Sclaroff, S.; and Saenko, K. 2020. Universal Domain Adaptation through Self Supervision. *Advances in Neural Information Processing Systems*, 33.

Saito, K.; Yamamoto, S.; Ushiku, Y.; and Harada, T. 2018. Open set domain adaptation by backpropagation. *In Proceedings of the European Conference on Computer Vision*, 153–168.

Tack, J.; Mo, S.; Jeong, J.; and Shin, J. 2020. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *arXiv preprint arXiv:2007.08176*.

Tang, H.; Chen, K.; and Jia, K. 2020. Unsupervised domain adaptation via structurally regularized deep clustering. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8725–8735.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep hashing network for unsupervised domain adaptation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 5018–5027.

Xu, R.; Li, G.; Yang, J.; and Lin, L. 2019. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. *In Proceedings of the*

IEEE/CVF International Conference on Computer Vision, 1426–1435.

Yan, H.; Ding, Y.; Li, P.; Wang, Q.; Xu, Y.; and Zuo, W. 2017. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2272–2281.

You, K.; Long, M.; Cao, Z.; Wang, J.; and Jordan, M. 2019. Universal domain adaptation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2720–2729.

Zhang, Y.; Liu, T.; Long, M.; and Jordan, M. 2019. Bridging theory and algorithm for domain adaptation. *In International Conference on Machine Learning*, 7404–7413.

Zhuang, C.; Zhai, A. L.; and Yamins, D. 2021. Local aggregation for unsupervised learning of visual embeddings. *In International Conference on Machine Learning*, 10738–10748.

Zong, B.; Song, Q.; Min, M. R.; Cheng, W.; Lumezanu, C.; Cho, D.; and Chen, H. 2018. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. *In International Conference on Learning Representations*.