Evidential Neighborhood Contrastive Learning for Universal Domain Adaptation

Liang Chen [1], **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 the knowledge learned from a labeled source domain to an unlabeled target domain without any constraints on the label sets. However, domain shift and category shift make UniDA extremely challenging, mainly attributed to the requirement of identifying both shared "known" samples and private "unknown" samples. Previous methods barely exploit the intrinsic manifold structure relationship between two domains for feature alignment, and they rely on the softmax-based scores with class competition nature to detect underlying "unknown" samples. Therefore, in this paper, we propose a novel eviden**T**ial **N**eighborhood con**T**rastive learning framework called TNT to address these issues. Specifically, TNT first proposes a new domain alignment principle: semantically consistent samples should be geometrically adjacent to each other, whether within or across domains. From this criterion, a cross domain multi-sample contrastive loss based on mutual nearest neighbors is designed to achieve common category matching and private category separation. Second, toward accurate "unknown" sample detection, TNT introduces a class competition-free uncertainty score from the perspective of evidential deep learning. Instead of setting a single threshold, TNT learns a category-aware heterogeneous threshold vector to reject diverse "unknown" samples. Extensive experiments on three benchmarks demonstrate that TNT significantly outperforms previous state-of-the-art UniDA methods.

Introduction

Deep neural network is data-hungry since it performs impressively on domains with abundant data labels, but it does not generalize well on new unlabeled domains. Task-related performance is significantly reduced owing to domain bias. Domain adaptation (DA) aims to solve this issue by eliminating feature discrepancy and transferring knowledge from the label-rich source domain to the label-scarce target domain (Ganin and Lempitsky 2015). Suppose L_s and L_t are the label sets in two domains, respectively. Traditional unsupervised DA usually assumes $L_s = L_t$, i.e., closed DA (CDA) (Tzeng et al. 2017). In complex real-world scenarios, however, this assumption may not be easily satisfied. Commonly, we may encounter $L_t \subset L_s$, i.e., partial DA (PDA) (Cao et al. 2018), or $L_s \subset L_t$, i.e., open-set DA (ODA) (Panareda and Gall 2017), or $L_s \cap L_t \neq \emptyset, L_s \cup L_t \neq L_s \, or \, L_t$, i.e., open partial DA (OPDA) (You et al. 2019). These variants have attracted the attention of the community in recent years and were resolved independently. However, negative aspect tends to confound this evolutionary process. Specifically, a method that is applicable to one variant may not be applicable to another variant. More realistically, we may not know in advance which of these variants will occur.

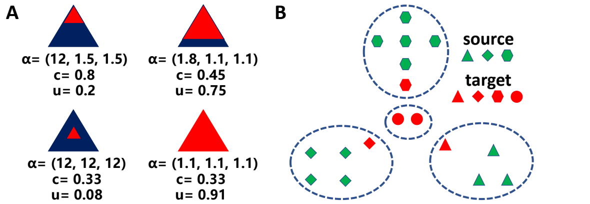


Figure 1: (A) Probability simplex with confidence (c) and uncertainty (u). The area of red indicates the uncertainty of prediction. Competition nature can produce high c but high u, or low c but low u. (B) Intra- and inter-domain mutual nearest neighbor principle ($k = 1$) for feature alignment.

Universal DA (UniDA) was proposed to account for both domain shift and category shift. It assumes that the two label spaces can be different and that their relationship is unknown in advance. In UniDA, we need to classify target samples into either one of the "known" labels or the "unknown" label. Here, however, UniDA poses two technical challenges. First, the removal of domain discrepancy should be constrained on the common categories between two domains, and we need to separate the respective private categories simultaneously. Second, in the absence of target label supervision, estimating the label distribution on the target domain and detecting potential target "unknown" samples is another main technical difficulty.

For the first challenge, UAN (You et al. 2019) and CMU (Fu et al. 2020) employ weighted adversarial network to discover shared label sets and promote common class adaptation. DANCE (Saito et al. 2020) uses a neighborhood clustering objective to move each target sample either to a source prototype or to its target neighbors. These methods hardly explore the manifold structural relationship between the two domains, and an explicit class-level feature alignment criterion is urgently required. For the second challenge, existing methods manually set a global threshold for softmax-based

[*] This work is completed in Huawei GoTen AI Lab.

[†] Corresponding Author.

confidence or entropy score to reject "unknown" samples. However, softmax loss introduces competition among different classes and can easily lead to over-confident predictions (see Figure 1(A)), thus making softmax-based uncertainty scores suboptimal for "unknown" sample detection. And as the "unknown" samples in nature belong to distinct semantic categories, their structural affinity with the source classes makes it difficult for a global threshold to divide diverse "unknown" samples.

In this paper, we address both of these challenges and propose an evidential neighborhood contrastive learning framework called TNT for UniDA. First, to enable the model to know "unknown", we formulate it as a Bayesian uncertainty estimation problem by introducing evidential deep learning (EDL) paradigm (Sensoy, Kaplan, and Kandemir 2018). EDL uses Multinomial-Dirichlet hierarchical model to predict the distribution of class probabilities, as well as provide the associated uncertainty. Based on it, a logarithmic evidence score, theoretically aligned with data likelihood, is proposed for "unknown" sample detection. Our mathematical insights and empirical results show that this uncertainty score is superior to the softmax-based score. To overcome the overfitting risk of EDL in a closed set, we propose an uncertainty versus confidence adversarial objective to calibrate the model prediction, which shapes the evidence surface and regularizes the evidence collection process. Instead of setting a global threshold, we learn a category-aware heterogeneous threshold vector to identify "unknown" samples more effectively.

Second, to match the common categories and separate respective private categories, we propose a novel domain alignment principle: semantically consistent samples should be geometrically adjacent to each other, whether within or across domains (see Figure 1(B)). Based on this criterion, we develop a neighborhood consensus contrastive learning paradigm to uncover the intrinsic manifold structure of both domains. Specifically, we first construct the intra- and inter-domain mutual nearest neighbor (MNN) pairs, which can be viewed as the positive pairs in the same category, otherwise negative pairs. Then a multi-sample contrastive loss is designed to integrate this knowledge of intra- and inter-domain positive and negative relations. To drive the domain alignment process, we minimize this contrastive loss to pull similar samples within and across domains closer and push dissimilar samples away to avoid negative label transfer.

Our contribution can be summarized as follows:

- We tackle the UniDA problem from a new perspective, i.e., performing Bayesian evidential learning and uncertainty estimation to support open-set knowledge transfer. An evidence-based uncertainty score with theoretical insight is introduced for "unknown" sample detection.
- We propose a novel feature alignment criterion for UniDA, i.e., that mutual nearest neighbors inside and across domains should be close to each other. By regarding them as a bridge for data integration, a cross-domain multi-sample contrastive loss is developed to remove domain bias and address category shift issues.
- We conduct extensive experiments on various UniDA benchmarks, and empirical results show that TNT outperforms previous state-of-the-art UniDA methods. Deeper analyses validate the effectiveness of the proposed uncertainty score and heterogeneous threshold.

Related Work

Universal domain adaptation

Universal DA is a challenging DA task, which assumes no prior knowledge about the relationship between source and target label spaces. You et al. (2019) proposed UAN to discover the shared classes between two domains by quantifying sample-level transferability. Fu et al. (2020) aggregated three complementary uncertainty measures, namely confidence, entropy and consistency, for accurate detection of target private classes. Saito et al. (2020) proposed DANCE to learn the target domain structure by neighborhood clustering, and used an entropy separation loss to achieve feature alignment. Li et al. (2021) proposed DCC that exploited domain consensus knowledge to discover discriminative clusters on both common and private samples. These methods do not consider the intrinsic manifold structure relationship between two domains, thus making them suboptimal for domain alignment. In this paper, we utilize intra- and inter-domain MNN pairs to bridge two domains.

Deep uncertainty learning

Understanding and quantifying uncertainty in neural network prediction is crucial for safe decision-making in high-risk fields (Gawlikowski et al. 2021). In recent years, researchers have shown an increased interest in uncertainty estimation in deep learning. Model-based and data-based uncertainty are two common ways to describe the predictive uncertainty of neural network (Choi et al. 2019; Bao, Yu, and Kong 2021). To distinguish the common and private categories between two domains, predictive uncertainty learned by deep neural networks can be a promising measure. Recently, evidential deep learning was developed to quantify classification uncertainty, which shows unprecedented success in the detection of out-of-distribution queries (Sensoy, Kaplan, and Kandemir 2018). In this paper, to the best of our knowledge, we are the first to incorporate an evidential learning module to differentiate "known" and potential "unknown" samples in UniDA.

Contrastive learning

Contrastive learning is a typical type of self-supervised learning paradigm (He et al. 2020; Chen et al. 2020). It learns representations by contrasting positive pairs against negative pairs. Many state-of-the-art methods for representation learning tasks are based on the contrastive learning framework (Chen and He 2021). Among them are instance-based (Grill et al. 2020), cluster-based (Caron et al. 2020), and neighbor-based contrastive learning techniques (Zhong et al. 2021). Although positive samples can come from augmented views of each instance, between-instance similarity conflicts with presumed instance distinction, impairing feature learning (Wang, Liu, and Yu 2021). In this article, we propose to utilize mutual nearest neighbors as positive pairs to achieve feature alignment between the two domains.

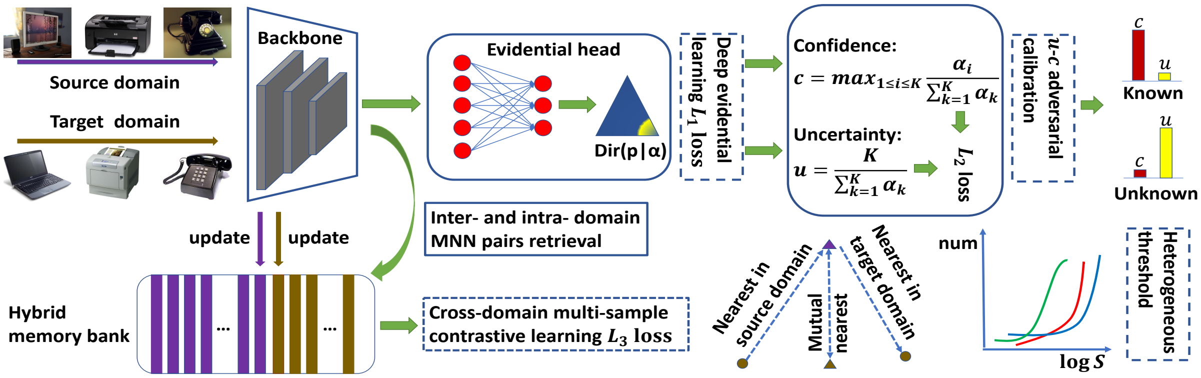

Backbone
Source domain
Target domain
Evidential head
Dir(p|α)
Deep evidential learning L_1 loss
Confidence:
$c = max_{1\le i\le K}\frac{\alpha_i}{\sum_{k=1}^{K}\alpha_k}$
Uncertainty:
$u = \frac{K}{\sum_{k=1}^{K}\alpha_k}$
L_2 loss
u-c adversarial calibration
c
u
Known
u
c
Unknown
update
update
Inter- and intra- domain MNN pairs retrieval
Hybrid memory bank
...
...
Cross-domain multi-sample contrastive learning L_3 loss
Nearest in source domain
Mutual nearest
Nearest in target domain
num
log S
Heterogeneous threshold

Figure 2: Schematics of TNT. Overall evidential learning model consists of a backbone and an evidential head. The cross-domain mutual nearest neighbor contrastive learning is proposed for domain alignment. The u-c adversarial mechanism is introduced for "unknown" sample detection. Category diversity leads to threshold heterogeneity.

Method

Overview

In UniDA, we have a labeled source domain $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ with L_s "known" categories, where $D_s \sim P_s$ along with an unlabeled target domain $D_t = \{(x_i^t)\}_{i=1}^{N_t}$ where $D_t \sim P_t$, and $P_s \neq P_t$. The target domain contains some "known" categories and potential "unknown" categories, and we denote its label space as L_t. We aim to learn a classification model and label the target samples with either one of the L_s "known" labels or the "unknown" label.

As shown in Figure 2, our model consists of two basic modules: (1) feature extractor g that maps input images into the embedding representations $z = g(x)$, and (2) evidential neural network head f which predicts the class-wise evidence corresponding to Dirichlet distribution parameters. The estimated evidence can further determine the predictive probabilities and uncertainty of the input. Then we employ the distribution characteristic of total evidences for "unknown" sample inference. To avoid model overfitting, we design an uncertainty versus confidence adversarial mechanism for prediction calibration. In training, we also propose a cross-domain mutual nearest neighbor contrastive learning module to drive the feature alignment process.

Deep evidential learning and uncertainty estimation

Existing DA models utilize a linear projection layer with a softmax operator on top of the deep neural network for discrimination on target data. The eventual model can be interpreted as a parameter regression framework of Multinomial distribution. In particular, for a L_s-class classification problem, assume discrete class probabilities $p = (p_1, p_2, ..., p_{L_s})$, as determined by network outputs; then the likelihood function of a labeled sample (x, y) is

$$\mathcal{L}(p|x) = Multinomial(y|p_1, p_2, ..., p_{L_s}) = \prod_{j=1}^{L_s} p_j^{y_j} \quad (1)$$

Minimizing the negative log-likelihood $-\log \mathcal{L}(p|x)$ over labeled samples with respect to network parameters is equivalent to cross-entropy loss. It only gives the point estimation of the Multinomial distribution over the categorical probabilities. Therefore, the output cannot capture the variance of predictive probabilities, i.e., second-order uncertainty. Besides, since the output probabilities have been squashed by the denominator of softmax, the network tends to produce an over-confident prediction for the "unknown" data (Wen et al. 2021). In ODA and OPDA settings, this phenomenon is even more common and detrimental.

To overcome the above limitations, evidential deep learning (EDL) formulates a principled way to jointly accomplish multi-class classification and uncertainty estimation by introducing the Bayesian hierarchical model. EDL introduces the Dirichlet distribution, a conjugate prior distribution of the Multinomial distribution, to represent the density of class probability assignment p. Specifically, assume that p follows a prior Dirichlet distribution with evidence parameter $\alpha = (\alpha_1, \alpha_2, ..., \alpha_{L_s}), \alpha_i > 1, \forall 1 \le i \le L_s$,

$$Dir(p|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{L_s} p_k^{\alpha_k - 1} \quad (2)$$

where $B(\alpha)$ is the Multinomial Beta function. Then training loss of the EDL model is the negative log-marginal likelihood, given by

$$\mathcal{L}_1 = \sum_{i=1}^{N_s} -\log\left(\int \prod_{j=1}^{L_s} p_{ij}^{y_{ij}^s} \frac{1}{B(\alpha_i^s)} \prod_{j=1}^{L_s} p_{ij}^{\alpha_{ij}^s - 1} dp_i\right) \quad (3)$$

$$= \sum_{i=1}^{N_s} \sum_{j=1}^{L_s} y_{ij}^s (\log S_i^s - \log \alpha_{ij}^s) \quad (4)$$

where S is the total evidence $S = \sum_{k=1}^{L_s} \alpha_k$. Here, α is the non-negative network prediction output and can be expressed as $\alpha = f(g(x)) + 1$. Based on the Dempster–Shafer Theory of Evidence (Sentz and Ferson 2002), the discriminative probability of k-th class is $p_k = \frac{\alpha_k}{S}$, and the prediction uncertainty u is inversely proportional to total evidence S, determined as $u = \frac{L_s}{S}$. In the training phase, by minimizing the $\mathcal{L}_1$ objective, we can collect the evidence for each known source category. Simultaneously, the obtained total evidence S or uncertainty u enables us to distinguish "known" from "unknown" samples in the inference process.

Discussion: Why is the total evidence score S more suitable than the softmax-based score for discovering potential target private classes? For ease of understanding, we assume that the activation function used by network f to predict α is an exponential function, and that the influence of constant 1 is ignored. Then we have

$$\log \max_i p(y_i|x) = \log \max_i \frac{\alpha_i}{S} = -\log S + \log \max_i \alpha_i \tag{5}$$

When we maximize $\log \max_i p(y_i|x)$ on labeled data, by minimizing the negative log-marginal likelihood $\mathcal{L}_1$, $\max_i \alpha_i$ tends to be higher and S tends to be lower. At this time, the prediction uncertainty u tends to be higher, which is beyond our expectations. For "known" categories, the prediction should be accurate and certain, while for "unknown" categories, the prediction uncertainty should be high, i.e., low S (Krishnan and Tickoo 2020). More importantly, as noted previously, the softmax-based score introduces competition among different classes, which can easily produce arbitrarily high score values, i.e., over-confident predictions. In contrast to the competition nature of softmax normalization, the total evidence score S is a statistic based on summation, free of any competition.

u-c adversarial mechanism for uncertainty calibration

Uncertainty characterizes the risk of differentiating the target sample into "known" categories. A natural way to discover potential target private categories is to set a threshold for the uncertainty measure. Samples with uncertainty larger than this threshold have high probabilities of being recognized as "unknown" label. However, minimizing $\mathcal{L}_1$ would enforce the total evidence S of "known" category samples to be compressed. Since uncertainty u is inversely proportional to S, minimizing $\mathcal{L}_1$ would have the potential risk that the uncertainty of the "known" category is large, and that the uncertainty of the "unknown" category is small, further causing the overfitting of the model. Meanwhile, the uncertainty gap between the "known" and "unknown" samples may not be optimal for discovering the underlying private categories.

We define $c = \max_k p_k$ as the prediction confidence. A well-calibrated prediction model should take into account both confidence and uncertainty. For "known" samples, confidence c is high, while uncertainty u is low, and for "unknown" samples, confidence c is low, while the uncertainty u is high. To strengthen the inverse relationship between them, we propose a u-c adversarial objective, defined as

$$\begin{aligned} \mathcal{L}_2 = \mathcal{L}_2^s + \mathcal{L}_2^t = &-\sum_{i=1}^{N_s} c_i^s \log(1-u_i^s) \\ &-\sum_{j=1}^{N_t} (c_j^t \log(1-u_j^t) + (1-c_j^t)\log u_j^t) \end{aligned} \tag{6}$$

The first term aims to give low uncertainty ($u \to 0$) and high confidence ($c \to 1$) on the labeled source domain, while the second term tries to penalize the u-versus-c homogeneity on the unlabeled target domain, forcing them to optimize in opposite directions. In this way, the model is encouraged to learn a skewed and sharp Dirichlet simplex for "known" categories and provide an unskewed and flat Dirichlet simplex for "unknown" samples. It can be seen that our u-c adversarial mechanism does not rely on additional validation set during training. Thus it provides better flexibility to calibrate the prediction uncertainty.

So far, we have not discussed how to determine a threshold δ to reject potential target private samples. Most previous methods use a validation set to set a single global threshold. Here we propose a category-aware threshold selection method based on logarithmic total evidence ($\log S$) distribution. The motivation behind it is that the diversity of "known" categories determines the heterogeneity of the threshold. Assume that the "unknown" threshold vector is $\hat{\delta} \in R^{L_s}$; then we label the target sample i by

$$y_i^t = \begin{cases} j, & \text{if } \log S_i^t \geq \hat{\delta}_j,\, j = \arg\max_{1 \leq k \leq L_s} \alpha_{ik}^t \\ unknown, & \text{if } \log S_i^t < \hat{\delta}_j,\, j = \arg\max_{1 \leq k \leq L_s} \alpha_{ik}^t \end{cases} \tag{7}$$

In particular, we first collect the logarithm total evidence score of the samples under each source category and record them as $\Omega_k = \{\log S_i^s : y_i^s = k\}, 1 \leq k \leq L_s$. Then we fit a Gaussian distribution on the Ω_k to obtain the mean estimation $\hat{\upsilon}_k$ and standard deviation estimation $\hat{\sigma}_k$. According to the "three-sigma" rule, we set $\hat{\delta}_k = \hat{\upsilon}_k - 2 \times \hat{\sigma}_k$, which can contain more than 95% source samples in each category and ignore minor outliers. Our threshold decision approach is data-based and learned from the source domain, which avoids tricky hyperparameter selection.

Mutual nearest neighbor contrastive learning for feature alignment

The interference of domain bias makes it difficult for a classifier trained on the source domain to obtain good generalization performance on the target domain. Therefore, eliminating domain discrepancy and learning domain invariant representation forms the basis for improving the generalization of the DA model. However, this goal is more difficult for UniDA, because we need to match common categories between the two domains, as well as separate the respective private categories. Alignment at the global domain level may reduce the feature margin between potential private classes and common classes, thereby compromising discrimination on the target domain. The dilemma of alignment at the class level is the category definition on the target domain together with the category relationship between the two domains. Geometric nearest neighbors are commonly used to describe similar patterns in manifold learning, so here we introduce a novel DA inductive bias: data points that are neighbors with each other are the mainstay of improving the compactness of each category in the domain, as well as the bridge of common category matching. On this basis, we propose a multi-sample contrastive learning paradigm that unites neighborhood consensus intra- and inter-domains.

Specifically, assuming that all embedding vectors $\{z_i^s\}_{i=1}^{N_s} \cup \{z_j^t\}_{j=1}^{N_t}$ are l_2 normalized, we first denote $N_k^s(i)$

Table 1: H-score comparison in OPDA setting. Some results are referred to previous work (Li et al. 2021).

Methods	Type	Office (10/10/11)							OfficeHome (10/5/50)													VisDA (6/3/3)
		A2W	D2W	W2D	A2D	D2A	W2A	Avg	A2C	A2P	A2R	C2A	C2P	C2R	P2A	P2C	P2R	R2A	R2C	R2P	Avg	S2R
RTN	C	50.2	54.7	55.2	50.2	47.7	49.3	51.2	38.4	44.7	45.7	42.6	44.1	45.5	42.6	36.8	45.5	44.6	39.8	44.5	42.9	26.0
IWAN	P	50.1	54.1	55.4	50.6	49.7	49.8	51.6	40.5	47.0	47.8	45.0	45.1	47.6	45.8	41.4	47.6	46.3	42.5	46.5	45.3	27.6
OSBP	O	50.2	55.5	57.2	51.1	49.8	50.2	52.3	39.6	45.1	46.2	45.7	45.2	46.8	45.3	40.5	45.8	45.1	41.6	46.9	44.5	27.3
UAN	U	58.6	70.6	71.4	59.7	60.1	60.3	63.5	51.6	51.7	54.3	61.7	57.6	61.9	50.4	47.6	61.5	62.9	52.6	65.2	56.6	30.5
CMU	U	67.3	79.3	80.4	68.1	71.4	72.2	73.1	56.0	56.9	59.2	67.0	64.3	67.8	54.7	51.1	66.4	68.2	57.9	69.7	61.6	34.6
DANCE	U	75.8	90.9	87.1	79.6	82.9	77.6	82.3	61.0	60.4	64.9	65.7	58.8	61.8	73.1	61.2	66.6	67.7	62.4	63.7	63.9	42.8
DCC	U	78.5	79.3	88.6	**88.5**	70.2	75.9	80.2	58.0	54.1	58.0	**74.6**	70.6	77.5	64.3	**73.6**	74.9	**81.0**	**75.1**	80.4	70.2	43.0
TNT	U	**80.4**	**92.0**	**91.2**	85.7	**83.8**	**79.1**	**85.4**	**61.9**	**74.6**	**80.2**	73.5	**71.4**	**79.6**	**74.2**	69.5	**82.7**	77.3	70.1	**81.2**	**74.7**	**55.3**

and $N_k^t(i)$ to represent the k nearest neighbor set measured by cosine distance of sample i in the source domain and target domain, respectively. If $i \in D_s$, its $N_k^s(i)$ can be replaced with those samples shared with consistent category label. Then the mutual nearest neighbor set $M_k^s(i)$ and $M_k^t(i)$ of sample i in the two domains can be defined as

$$M_k^s(i) = \begin{cases} \{j \in D_s : y_i = y_j\}, & \text{if } i \in D_s \\ \{j \in D_s : i \in N_k^t(j) \cap j \in N_k^s(i)\}, & \text{if } i \in D_t \end{cases} \tag{8}$$

$$M_k^t(i) = \begin{cases} \{l \in D_t : i \in N_k^s(l) \cap l \in N_k^t(i)\}, & \text{if } i \in D_s \\ \{l \in D_t : i \in N_k^t(l) \cap l \in N_k^t(i)\}, & \text{if } i \in D_t \end{cases} \tag{9}$$

Here we aim to pull mutual nearest neighbor pairs across the source and target domains closer to each other, while pushing away those non-geometrically close samples. Inspired by InfoNCE loss (Hjelm et al. 2018), we propose a novel cross-domain multi-sample contrastive learning objective function, given by

$$\mathcal{L}_c^i = \begin{cases} \text{if } i \in D_s, \\ -\log \frac{\sum_{j \in M_k^s(i)} \exp(z_i^s z_j^s/\tau) + \sum_{l \in M_k^t(i)} \exp(z_i^s z_l^t/\tau)}{\sum_{m=1}^{N_s} \exp(z_i^s z_m^s/\tau) + \sum_{n=1}^{N_t} \exp(z_i^s z_n^t/\tau)}; \\ \text{if } i \in D_t, \\ -\log \frac{\sum_{j \in M_k^s(i)} \exp(z_i^t z_j^s/\tau) + \sum_{l \in M_k^t(i)} \exp(z_i^t z_l^t/\tau)}{\sum_{m=1}^{N_s} \exp(z_i^t z_m^s/\tau) + \sum_{n=1}^{N_t} \exp(z_i^t z_n^t/\tau)}. \end{cases} \tag{10}$$

where τ is the temperature parameter. In training, we split source and target samples into different mini-batches and forward them separately. Let B_s and B_t denote the source and target batches respectively, then the overall training contrastive loss is computed as the sum across all the source samples from B_s and target samples from B_t,

$$\mathcal{L}_3 = \sum_{i \in B_s} \mathcal{L}_c^i + \sum_{j \in B_t} \mathcal{L}_c^j \tag{11}$$

As defined above, the neighbor identification procedures and the loss function implicitly cover computations involving all the embedded features of two domains, which soon become intractable for large datasets. To address this issue, we employ a hybrid memory bank $\bar{Z} = \{\bar{z}_1^s, ..., \bar{z}_{N_s}^s, \bar{z}_1^t, ..., \bar{z}_{N_t}^t\}$ to maintain the running average of all source and target features. We initialize the memory bank with random unit vectors and update its values by mixing $\bar{z}_i$ and z_i during training as follows, where γ is a mixing parameter,

$$\bar{z}_i \leftarrow \gamma \bar{z}_i + (1-\gamma) z_i \tag{12}$$

Overall objective. The model is jointly optimized with three terms, i.e., evidential deep learning loss $\mathcal{L}_1$, uncertainty calibration loss $\mathcal{L}_2$ and contrastive feature alignment loss $\mathcal{L}_3$,

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \lambda \mathcal{L}_3 \tag{13}$$

where λ is set as 0.1 to balance each loss component.

Experiment

Setup

Dataset. We conduct experiments on three benchmark datasets. **Office** (Saenko et al. 2010) consists of about 4700 images in 31 categories from three domains: Amazon (A), DSLR (D), and Webcam (W). **OfficeHome** (Venkateswara et al. 2017) is a larger dataset with 15500 images from 65 categories in four domains: Artistic images (A), Clip-Art images (C), Product images (P), and Real-World images (R). **VisDA** (Peng et al. 2017) is a large-scale challenging dataset with 12 categories, with source domain containing about 150K synthetic images (S) and target domain containing 50K real world images (R). Let $|L_s \cap L_t|$, $|L_s - L_t|$ and $|L_t - L_s|$ denote the number of common categories, source private categories and target private categories, respectively. Following existing studies, we show the category split ($|L_s \cap L_t|/||L_s - L_t||/|L_t - L_s|$) of each experimental setting in a corresponding result table. The split details can be seen in the supplemental material.

Evaluation protocols. We use the same evaluation metrics as those in the previous study (Fu et al. 2020). In CDA and PDA settings, we calculate the classification accuracy over all target samples. In ODA and OPDA settings, target private samples are grouped into a single "unknown" class. As such, the trade-off between the accuracy of "known" and "unknown" classes is important in evaluating performance. Thus, we use the H-score, i.e. the harmonic mean of the accuracy on common classes and accuracy on "unknown" class, to evaluate each method. The H-score metric is high only when both the "known" and "unknown" accuracies are high. For all experiments, the averaged results of three runs are reported. Additionally, we assume no prior information about category shift in any of the above DA settings.

Implementation details. Our implementation is based on PyTorch and we conduct all experiments on one Tesla V100 GPU. The network backbone is ResNet50 (He et al. 2016) pretrained on ImageNet (Deng et al. 2009), and the evidential head consists of two fully-connected layers. In the training phase, we choose the exp function as the evidence function, because we empirically found it to be numerically more stable when training the evidential loss $\mathcal{L}_1$. Following previous work (Saito et al. 2020), the batch size is set to 36 and the temperature parameter τ is set as 0.05. The memory bank is updated with momentum $\gamma = 0.5$ and the nearest neighbor number k is set to 30 for Office and OfficeHome and 50 for

Table 2: H-score comparison in ODA setting. Some results are referred to previous work (Li et al. 2021).

		Office (10/0/11)							**OfficeHome (25/0/40)**													**VisDA (6/0/6)**
Methods	**Type**	A2W	A2D	D2W	W2D	D2A	W2A	Avg	A2C	A2P	A2R	C2A	C2P	C2R	P2A	P2C	P2R	R2A	R2C	R2P	Avg	S2R
STA	O	75.9	75.0	69.8	75.2	73.2	66.1	72.5	55.8	54.0	68.3	57.4	60.4	66.8	61.9	53.2	69.5	67.1	54.5	64.5	61.1	-
OSBP	O	**82.7**	82.4	**97.2**	91.1	75.1	73.7	83.7	55.1	65.2	72.9	64.3	64.7	**70.6**	63.2	53.2	73.9	66.7	54.5	72.3	64.7	52.3
ROS	O	82.1	82.4	96.0	**99.7**	77.9	77.2	85.9	60.1	**69.3**	**76.5**	58.9	65.2	68.6	60.6	56.3	**74.4**	**68.8**	60.4	**75.7**	66.2	-
UAN	U	46.8	38.9	68.8	53.0	68.0	54.9	55.1	0.0	0.0	0.2	0.0	0.2	0.2	0.0	0.0	0.2	0.2	0.0	0.1	0.1	51.9
CMU	U	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	54.2
DANCE	U	78.8	84.9	78.8	88.9	79.1	68.3	79.8	61.9	61.3	63.7	64.2	58.6	62.6	**67.4**	**61.0**	65.5	65.9	61.3	64.2	63.0	67.5
DCC	U	54.8	58.3	89.4	80.9	67.2	**85.3**	72.6	56.1	67.5	66.7	49.6	66.5	64.0	55.8	53.0	70.5	61.6	57.2	71.9	61.7	59.6
TNT	U	82.3	**85.8**	91.2	96.2	**80.7**	81.5	**86.3**	**63.4**	67.9	74.9	**65.7**	**67.1**	68.3	64.5	58.1	73.2	67.8	**61.9**	74.5	**67.3**	**71.6**

Table 3: Accuracy comparison in PDA setting. Some results are referred to previous works (Saito et al. 2020; Li et al. 2021).

		Office (10/21/0)							**OfficeHome (25/40/0)**													**VisDA (6/6/0)**
Methods	**Type**	A2W	A2D	D2W	W2D	D2A	W2A	Avg	A2C	A2P	A2R	C2A	C2P	C2R	P2A	P2C	P2R	R2A	R2C	R2P	Avg	S2R
PADA	P	82.2	86.5	92.7	99.3	95.4	**100.0**	92.7	52.0	67.0	78.7	52.2	53.8	59.1	52.6	43.2	78.8	73.7	56.6	77.1	62.1	-
ETN	P	94.5	95.0	**100.0**	**100.0**	**96.2**	94.6	96.7	59.2	77.0	79.5	62.9	65.7	75.0	68.3	55.4	84.4	75.7	57.7	84.5	70.5	59.8
BA^{3}US	P	**98.9**	**99.4**	**100.0**	98.7	94.8	95.0	**97.8**	**60.6**	**83.2**	**88.4**	71.8	**72.8**	83.4	**75.5**	61.6	**86.5**	79.3	62.8	**86.1**	**76.0**	54.9
UAN	U	76.8	79.7	93.4	98.3	82.7	83.7	85.8	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	U	84.2	84.1	97.2	98.8	69.2	66.8	83.4	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	U	71.2	77.1	94.6	96.8	83.7	92.6	86.0	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	U	81.3	87.3	**100.0**	**100.0**	95.4	95.5	93.3	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
TNT	U	83.4	88.2	98.5	98.6	92.7	93.9	92.5	55.1	75.3	84.6	72.9	70.0	82.5	71.4	58.7	83.3	79.1	62.4	83.2	73.2	**75.2**

VisDA as default. We train our model for 10000 iterations with Nestrov momentum SGD. The initial learning rate is set to 0.001, which is decayed with the same schedule as in previous studies (Long et al. 2018; Saito et al. 2020).

Results Comparison

Comparison baselines. We compare TNT with previous state-of-the-arts in four possible scenarios of UniDA, i.e., CDA (RTN (Long et al. 2016), CDAN (Long et al. 2018), MDD (Li et al. 2020), SRDC (Tang, Chen, and Jia 2020)), PDA (PADA (Cao et al. 2018), IWAN (Zhang et al. 2018), ETN (Cao et al. 2019), BA^{3}US (Liang et al. 2020)), ODA (OSBP (Saito et al. 2018), STA (Liu et al. 2019), ROS (Bucci, Loghmani, and Tommasi 2020)) and OPDA (UAN (You et al. 2019), CMU (Fu et al. 2020), DANCE (Saito et al. 2020), DCC (Li et al. 2021)). For all cases, each UniDA method is tested without knowing the prior of category shift, and those baselines tailed for each setting are conducted by taking this prior into consideration. We use "C", "P", "O" and "U" to denote the methods specifically designed for CDA, PDA, ODA and UniDA accordingly. Due to a limited space, we put some results in supplementary.

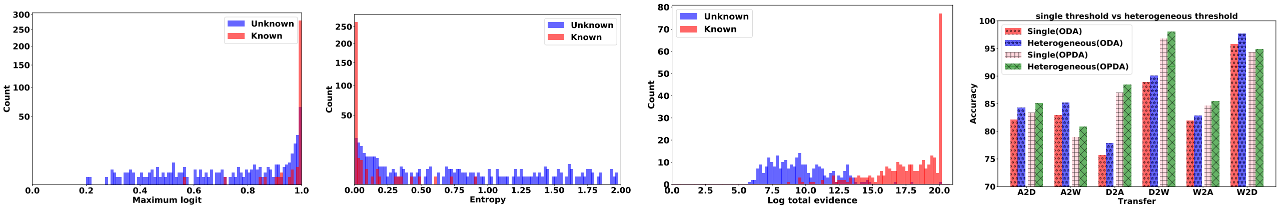
ODA and OPDA settings. From the results in Table 1, TNT achieves a new state-of-the-art on three datasets in the most challenging OPDA setting. With respect to H-score, TNT outperforms DANCE on Office by 3% and DCC on OfficeHome by 4%. On the large-scale VisDA dataset, TNT gives more than 10% improvement compared to all other methods in terms of H-score. Collectively, this evidence shows that TNT gains a better trade-off between common categories classification and private samples identification. For the ODA setting, the H-score comparison results are presented in Table 2. TNT consistently performs better than all UniDA baselines on three benchmarks, with +4% H-score improvement. Even compared with ROS, a previous state-of-the-art method tailed for ODA setting, TNT is also slightly superior on Office and OfficeHome datasets. Under these two scenarios with "unknown" samples, our method shows a stronger capability on the separation of common and private categories, which benefits from the MNN contrastive feature alignment and uncertainty calibration.

CDA and PDA settings. In the PDA setting, the results in Table 3 tell us that TNT outperforms all other baselines including those tailored for PDA on VisDA. For Office and OfficeHome datasets, TNT also gives comparable results to BA^{3}US, which is one of the state-of-the-art methods in PDA. The results in the CDA setting show that TNT outperforms other state-of-the-art UniDA methods on three datasets (see supplementary). Even compared to those methods specialized in the CDA setting, TNT achieves comparable performance to some of them, such as only inferior to SRDC on Office and OfficeHome and to MDD on VisDA. However, such methods customized for these two settings cannot adapt to situations where "unknown" samples exist, thereby limiting their application in real-world scenarios.

Total evidence score works better than softmax-based score in "known" and "unknown" separation. We begin by assessing the improvement of total evidence score over softmax-based scores. Figure 3 compares the histogram distributions of total evidence score, softmax-based confidence score and softmax-based entropy score. The evidence scores naturally form smooth bimodal distribution to separate "known" and "unknown" samples clearly. In contrast, confidence score and entropy score fail to distinguish them obviously, because many "unknown" samples also give high confidence and low entropy. To gain further quantitative insights, we calculate the AUC value which measures how well "known" and "unknown" samples are separated. The performance of total evidence score consistently outperforms the confidence score and entropy score by a large margin. Overall our experiment shows that the proposed total evidence score enables more effective "unknown" sample detection and as a result, is a promising anomaly measure.

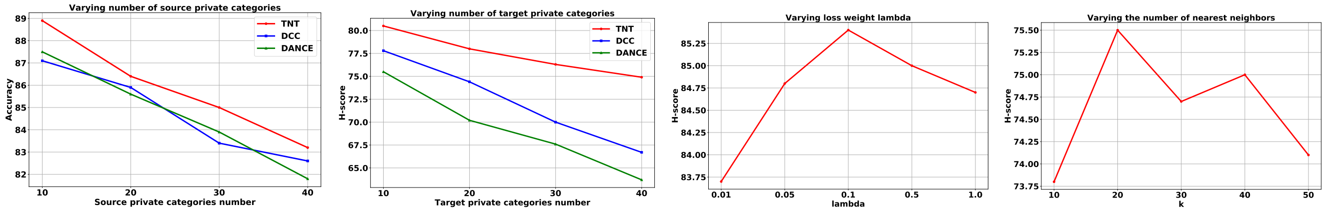
The heterogeneous threshold is superior to the single global threshold. We analyze the impact of the category-aware heterogeneous threshold on the identification of "unknown" samples. The ODA and OPDA settings on Office were used for this experiment. We report the "unknown" sample detection accuracy in Figure 3(d), where a single global threshold is obtained by using the mean and standard deviation of all source data. Whether ODA or OPDA, we can see that the accuracies under the heterogeneous threshold are higher than the results under the single global threshold, fully confirming the need to consider threshold heterogeneity caused by the diversity of categories.

Number of source private categories. We compare the behavior of TNT with DANCE and DCC under different



(a) Confidence (AUC=0.825) (b) Entropy (AUC=0.834) (c) Total evidence (AUC=0.976) (d) Threshold comparison

Figure 3: (a, b, c) Histogram of three uncertainty scores on "D2W" of Office in ODA setting. The AUC value of each score is also given. (d) Performance comparison between global and heterogeneous threshold on Office in ODA and OPDA setting.



(a) Real-World to Product (b) Art to Real-World (c) Loss weight λ (d) Neighbors number k

Figure 4: Various case studies, including source and target private classes number, loss weight and nearest neighbor number.

number of source private categories in the PDA setting. In this analysis, we use "R2P" in OfficeHome to conduct experiments, where there are 25 common categories between two domains.We vary the class number present only in the source domain from 10 to 40. The accuracy result is shown in Figure 4(a). With the appearance of more unshared private categories in the source domain, the performance of the three methods degrades. However, TNT consistently outperforms DANCE and DCC, indicating that it is robust to the change of source private class number.

Number of target private categories. We also analyze the behavior of TNT under the different "unknown" classes number. Here we perform ODA experiment on "A2R" task in OfficeHome, which has 25 shared classes. We increase the number of "unknown" classes only in target domain from 10 to 40. Figure 4(b) shows the H-score comparison among three methods. As we add more target private categories, the H-score of all methods decreases. However, TNT consistently performs better than DANCE and DCC, validating its stability with respect to the "unknown" classes number.

Feature visualization. We use t-SNE to visualize the learned target features with corresponding ground-truth labels and our predicted labels in Figure 5, under the ODA setting. Most target features between "known" common categories and "unknown" private categories are well separated, and those samples in the same class are grouped together. This mainly results from our feature alignment strategy, i.e., mutual nearest neighbor contrastive learning.

Ablation Study

Effect of $\mathcal{L}_2$, $\mathcal{L}_3$. To evaluate the contribution of $\mathcal{L}_2$ and $\mathcal{L}_3$, we train the model with $\mathcal{L}_1$ and each component alone. In this analysis, we use the VisDA dataset on four DA settings and present the results in Table 4. It can be seen that removing either $\mathcal{L}_2$ or $\mathcal{L}_3$ would serve to degrade the performance. The effect of $\mathcal{L}_3$ is more significant, since $\mathcal{L}_3$ aims to remove the domain discrepancy and achieve common category matching and private category separation. Without $\mathcal{L}_2$ for prediction calibration, the classification boundary would be fuzzy, thereby harming accuracy.

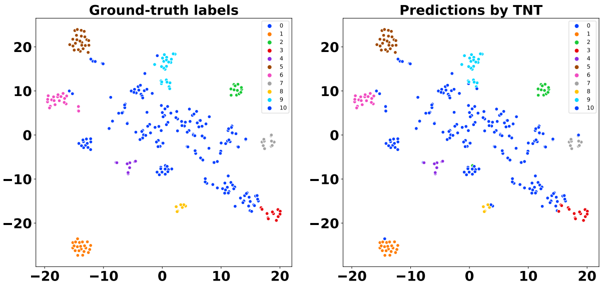


Figure 5: Feature visualization on "W2D" in Office. Blue plots are "unknown" samples, others are "known" samples.

Hyperparameter sensitivity. To show the sensitivity of TNT to the loss weight λ, we conducted control experiments on Office under the OPDA setting, and the results are presented in Figure 4(c). Within a wide range of $\lambda \in [0.01, 1.0]$, the performance changes very little, showing that TNT is robust to the selection of λ. We also analyze the behavior of TNT when changing the nearest neighbor number k on OfficeHome in the OPDA setting. As shown in Figure 4(d), the H-score only varies slightly with $k \in [10, 50]$, validating that TNT is stable to the choices of k.

Table 4: Ablation study on VisDA dataset in four settings.

VisDA	CDA	PDA	ODA	OPDA
TNT w/o $\mathcal{L}_2$	69.5	72.6	66.8	50.4
TNT w/o $\mathcal{L}_3$	64.7	65.1	63.5	46.9
TNT (full)	72.3	75.2	71.6	55.3

Conclusion

In this paper, we introduce a novel UniDA framework called TNT from the perspective of evidential deep learning and contrastive learning. It consists of two effective modules: contrastive feature alignment based on intra- and inter-domain mutual nearest neighbor pairs, an evidence-based uncertainty score together with a category-aware heterogeneous threshold vector for "unknown" sample detection. A thorough evaluation on three benchmarks shows the superior performance of TNT, compared to previous state-of-the-arts.

References

Bao, W.; Yu, Q.; and Kong, Y. 2021. Evidential Deep Learning for Open Set Action Recognition. *arXiv preprint arXiv:2107.10161*.

Bucci, S.; Loghmani, M. R.; and Tommasi, T. 2020. On the effectiveness of image rotation for open set domain adaptation. *In European Conference on Computer Vision*, 422–438.

Cao, Z.; Ma, L.; Long, M.; and Wang, J. 2018. Partial adversarial domain adaptation. *In Proceedings of the European Conference on Computer Vision*, 135–150.

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 Conference on Computer Vision and Pattern Recognition*, 2985–2994.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *In Thirty-fourth Conference on Neural Information Processing Systems*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. *In International conference on machine learning*, 1597–1607.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 15750–15758.

Choi, J.; Chun, D.; Kim, H.; and Lee, H.-J. 2019. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. *In Proceedings of the IEEE International Conference on Computer Vision*, 502–511.

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

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.

Gawlikowski, J.; Rovile, C.; Tassi, N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; and Kruspe, A. e. a. 2021. A Survey of Uncertainty in Deep Neural Networks. *arXiv preprint arXiv:2107.03342*.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

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.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Krishnan, R.; and Tickoo, O. 2020. Improving model calibration with accuracy versus uncertainty optimization. *Advances in Neural Information Processing Systems*, 33.

Li, G.; Kang, G.; Zhu, Y.; Wei, Y.; and Yang, Y. 2021. Domain Consensus Clustering for Universal Domain Adaptation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9757–9766.

Li, J.; Chen, E.; Ding, Z.; Zhu, L.; Lu, K.; and Shen, H. 2020. Maximum density divergence for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*.

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. W.; and Yang, Q. 2019. Separate to adapt: Open set domain adaptation via progressive separation. *In Proceedings of the IEEE 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.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2016. Unsupervised Domain Adaptation with Residual Transfer Networks. *Advances in Neural Information Processing Systems*, 136–144.

Panareda, P.; and Gall, J. 2017. Open set domain adaptation. *In Proceedings of the IEEE International Conference on Computer Vision*, 754–763.

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*.

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.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *In Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 3183–3193.

Sentz, K.; and Ferson, S. 2002. Combination of evidence in Dempster-Shafer theory. *Albuquerque: Sandia National Laboratories*, 4015.

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

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 7167–7176.

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.

Wang, X.; Liu, Z.; and Yu, S. X. 2021. Unsupervised Feature Learning by Cross-Level Instance-Group Discrimination. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12586–12595.

Wen, Y.; Liu, W.; Weller, A.; Raj, B.; and Singh, R. 2021. SphereFace2: Binary Classification is All You Need for Deep Face Recognition. *arXiv preprint arXiv:2108.01513*.

You, K.; Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2019. Universal domain adaptation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2720–2729.

Zhang, J.; Ding, Z.; Li, W.; and Ogunbona, P. 2018. Importance weighted adversarial nets for partial domain adaptation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 8156–8164.

Zhong, Z.; Fini, E.; Roy, S.; Luo, Z.; Ricci, E.; and Sebe, N. 2021. Neighborhood Contrastive Learning for Novel Class Discovery. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10867–10875.