A Novel Framework based on Unknown Estimation via Principal Sub-space for Universal Domain Adaption

Yifan Wang, Lin Zhang, Ran Song, Lin Ma, Wei Zhang

Abstract—Universal domain adaptation (UniDA) aims to transfer the knowledge of common classes from source domain to target domain without any prior knowledge on the label set, which requires to distinguish the unknown samples from the known ones in the target domain. Like the traditional unsupervised domain adaptation problem, the misalignment between two domains exists due to the biased and less-discriminative embedding. Recent methods proposed to complete the domain misalignment by clustering target samples with the nearest neighbors or the prototypes. However, it is dangerous to do so since we do not have any prior knowledge about the distributions of unknown samples which can magnify the misalignment especially when the unknown set is big. Meanwhile, other existing classifier-based methods could easily produce overconfident predictions of unknown samples because of the supervised objective in source domain leading the whole model to be biased towards the common classes in the target domain. Therefore, to deal with two issues above, this paper proposed a novel UniDA framework. First, we infer an empirical estimation value of the posterior probability of a target sample belonging to the unknown class based on its neighborhood searched from source domain. Then, we propose a novel non-parameter unknown samples detection method based on mapping the samples in the original feature space into a reliable linear sub-space which makes data points more sparse to reduce the misalignment between unknown samples and source samples. Moreover, unlike the recent methods applying extra parameters to improve the classification of unknown samples, this paper well balances the confidence values of both known and unknown samples through an unknown-adaptive margin loss which can control the gradient updating of the classifier learning on supervised source samples depending on the confidence level of detected unknown samples at current step. Finally, experiments on four public datasets demonstrate that our method significantly outperforms existing state-of-the-art methods.

1 INTRODUCTION

Unsupervised domain adaptation (UDA) [1], [2], [3], [4], [5] aims to transfer the learned knowledge from the labeled source domain to the unlabeled target domain so that the inter-sample affinities in the latter can be properly measured. The assumption of traditional unsupervised DA, i.e. closed-set DA, is that the source domain shares identical label set with the target domain, which significantly limits its applications in real-world scenarios. Thus, relaxations to this assumption have been investigated. Partial-set DA (PDA) [6], [7], [8], [9] assumes that target domain is not identical to source domain but its subset. On the contrary, Open-set DA (ODA) [10], [11], [12] assumes that target domain contains classes unknown to the source domain such that the source domain is a subset of the target domain. Open-partial DA (OPDA) [13], [14], [15] introduces private classes in both domains respectively, where the private classes in the target domain are unknown, and assumes that the common classes shared by the two domains have been identified. Universal DA (UniDA) [13], [14], [16] concerns about unsupervised DA with the most general setting, where no prior knowledge is required on the label set relationship between domains.

Similar to the traditional UDA problem, the domain misalignment exits due to the biased and less-discriminative embedding which can mislead the knowledge transformation and lead to the incorrect classification. In UniDA problem, the label spaces of two domains are not exactly overlapped which magnifies the domain bias. Thus, besides the misalignment between source and target common samples, a main challenge of UniDA is distinguishing the unknown target samples and avoiding mismatching the target samples to the source private classes. To deal with that, a popular method [13], [15], [16], [17] for UniDA is to employ a classifier which produces a confidence of each target sample or using the entropy of the prediction output by the classifier to determine whether it belongs to a particular known class seen in the source domain or the unknown class. They usually train their classifier on the source domain to generate confident predictions for target samples from common classes. However, the supervised objective in source domain can lead the whole model to be biased towards the common classes in the target domain, which can make the predictions of many unknown samples overconfident leading to the wrong classification and as mentioned by Chen et al. [18], the class competition nature may also cause the neural network to generate overconfident predictions for unknown instances. To deal with that, some recent approaches applied extra parameters to help classify the unknown samples, like Fu et al. [15] employed multiple classifiers to detect the target “unknown” samples by a mixture of uncertainties, Saito et al. [13] proposed to use an One-vs-All classifier to distinguish the unknown samples and Chen et al. [18] extended the softmax-based classifier to produce an energy-based uncertainty to determine the unknown samples.

Another kind of popular method [14], [18], [19] is to...
complete the alignment between samples in common classes of both source and target domains and push the unknown samples away from common classes. For instance, Saito et al. [19] proposed a prototype-based method to move each target sample either to a prototype of a source class or to its target neighbors. Li et al. [18] solved this problem by replacing the classifier-based framework with a clustering-based one which exploited the intrinsic structure of samples and thus increased the inter-sample affinity in each cluster. Chen et al. [18] proposed a geometric anchor-guided adversarial and contrastive learning framework with uncertainty modeling and achieved the state-of-the-art (SOTA) by making samples from each domain closer to their neighbors in both source and target domains.

However, without any prior knowledge about unknown samples and source private classes, approaches of completing the alignment between two domains are dangerous which can even magnify the misalignment. Since we have no idea about the diversity of semantic in the unknown set, compared to the inter-sample affinity in a known object class, that in the unknown class can be semantically much larger as the unknown class can contain samples that semantically belong to different object classes especially when the unknown set is large. This means that the inter-sample affinity between two samples in the unknown class could be even lower than that between a sample in the unknown class and a sample in a known object class, and due to the less-discriminative embedding, the inter-sample affinities between a known sample and unknown samples can be greater than that between the known sample and samples in the corresponding source class. As a result, some unknown samples could be easily pushed closer to one of the source classes incorrectly and some known samples can be clustered with the unknown samples which can aggravate the domain misalignment. As illustrated in Fig. 1, unknown samples with ‘Pan’ object class can be easily classified as ‘Marker’ or ‘Pencil’ which are similar to ‘Pan’ semantically. And due to the difference between images from different domains can cause the false classification on target common samples with ‘Marker’ class and the distribution of some unknown samples like ‘Bed’ images can be too close to the ‘Eraser’ source class even without high similarity on semantic. Thus, it is hard to complete the domain alignment without any prior knowledge about the distribution of unknown samples. Therefore, to deal with the above two issues, we propose a novel UniDA framework to reduce the influence of the domain misalignment and balance the confidences of known and unknown samples to deliver a better classification performance.

First, to distinguish the unknown samples from the target samples, we infer an empirical estimation of the posterior probability of a target sample being an unknown sample through its neighborhood searched from the source domain. Based on the empirical estimation, we observe that the biggest number of neighbors belonging to the same class is the key factor to distinguish the unknown samples and is negative correlated with the probability of a target sample being unknown. Next, since the precision of the empirical estimation is influenced by the domain misalignment, we propose a novel non-parametric unknown detection scheme by finding a linear sub-space of the original feature space where the distribution of both source and target samples is more sparse and discriminative. To achieve that, we analyse the covariance matrix of all source samples and the target samples in a batch. First, since the less-discriminative embedding can lead to the distribution of target samples biasing to that of source samples which can result in the misalignment between unknown target samples and source samples, we decrease the covariance of two samples to reduce the correlation of them and make data points more sparse, which can reduce the information redundancy of each sample caused by the less-discriminative embedding and adapt the biased distribution of target samples. Next, to avoid producing new misalignment due to the information loss after the feature dimension reduction, we maximum the variance of each sample to find the major dimensions which can maintain the feature of it as much as possible and cut off the minor dimensions. Then, we can adapt the wrong alignments between two samples and get a reliable linear sub-space.

In result, by searching the neighbors of each target samples from the source domain in the sub-space. We can estimate the probability of a target sample being unknown. Moreover, since the classifier learning on supervised source samples can lead to an imbalance problem of the predictions of target samples where the classifier can be biased to the source classes and generate the overconfident predictions for unknown samples, we propose a novel unknown-adaptive margin loss (UAM) to control the gradient updating rate of the classifier to avoid the prediction of a target sample converging into a particular source class where the margin depends on the current confidence level of detected unknown samples. A Kullback-Leibler divergence loss is employed to generate an uniform class posterior distribution for unknown samples found by the unknown detection to increase the entropy of the prediction. Notably, we do not introduce extra parameters to help the classification on
unknown samples like the proposed methods usually do [7], [9], [13].

- We are the first to give an empirical estimation of posterior probability of a target sample being an unknown sample through its neighborhood information searched from the source domain in UniDA. Base on the empirical estimation, we propose to use the biggest number of neighbors belonging to the same class searched from the source domain to determine a target sample is unknown or not.
- Based on the estimation of the probabilities of target samples belonging to the unknown class, we propose a novel non-parameter unknown detection method via mapping the features in the original feature space into a linear sub-space where the distribution of samples in the sub-space is more sparse and discriminative which can reduce the domain misalignment.
- We propose a novel unknown-adaptive margin loss to balance confidences of known target samples and unknown target samples without applying extra parameters. The margin loss can adeptly control the gradient updating of the classifier learning on the source domain based on the mean entropy output by the classifier of detected unknown samples currently, which can avoid the overconfident predictions of unknown samples while it can produce predictions with high confidence.

## 2 Related Work

We briefly review recent unsupervised DA methods in this section. According to the assumption made about the relationship between the label sets of different domains, we group these methods into three categories, namely PDA, ODA and UDA. We also briefly review a related problem named Out-of-Distribution detection to illustrate the relationship with our work.

### 2.1 Partial-set Domain Adaptation

PDA requires that the source label set is larger than and contains the target label set. Many methods for PDA have been developed [6], [7], [8], [9], [20]. For example, Cao et al. [20] presented the selective adversarial network (SAN), which simultaneously circumvented negative transfer and promoted positive transfer to align the distributions of samples in a fine-grained manner. Zhang et al. [8] proposed to identify common samples associated with domain similarities from the domain discriminator, and conducted a weighting operation based on such similarities for adversarial training. Cao et al. [7] proposed a progressive weighting scheme to estimate the transferability of source samples. Liang et al. [9] introduced balanced adversarial alignment and adaptive uncertainty suppression to avoid negative transfer and uncertainty propagation.

### 2.2 Open-set Domain Adaptation

ODA, first introduced by Busto et al. [10], assumes that there are private and common classes in both source and target domains, and common class labels are known as prior knowledge. They introduced the Assign-and-Transform-Iteratively (ATI) algorithm to address this challenging problem. Recently, one of the most popular strategies for ODA is to draw the knowledge from the domain discriminator to identify common samples across domains. Saito et al. [11] proposed an adversarial learning framework to obtain a boundary between source and target samples whereas the feature generator was trained to make the target samples far from the boundary. Bucci et al. [16] employed self-supervised learning technique to achieve the known/unknown separation and domain alignment.

### 2.3 Universal Domain Adaptation

UniDA, first introduced by You et al. [17] is subject to the most general setting of unsupervised DA, which involves no prior knowledge about the difference on object classes between the two domains. You et al. also presented an universal adaptation network (UAN) to evaluate the transferability of samples based on uncertainty and domain similarity for solving the UDA problem. However, the uncertainty and domain similarity measurements are sometimes not robust and sufficiently discriminative. Thus, Fu et al. [15] proposed another transferability measure, known as Calibrated Multiple Uncertainties (CMU), estimated by a mixture of uncertainties which accurately quantified the inclination of a target sample to the common classes. Li et al. [14] introduced Domain Consensus Clustering (DCC) to exploit the domain consensus knowledge for discovering discriminative clusters on the samples, which differed the unknown classes from the common ones. OVANet [13], proposed by Saito et al., trained a one-vs-all classifier for each class using labeled source samples and adapted the open-set classifier to the target domain. Recently, Chen et al. [18] proposed a geometric anchor-guided adversarial and contrastive learning framework with uncertainty modeling and achieve the state-of-the-art (SOTA) by exploring a new neighbors clustering method to complete the domain alignment and extend the traditional softmax-based classifier to the energy-based classifier. However, all recent method ignore that adapting the domain misalignment between two domains are dangerous since we do not have any knowledge about the source private classes and the unknown target samples. Especially, they could not perform well in the scenario of the unknown set being large.

### 2.4 Out-of-Distribution Detection

Out-of-Distribution (OOD) detection aims to detect OOD samples during the inference process which is similar to the UniDA problem of detecting the unknown samples. A baseline method by Hendrycks et al. [22] was proposed for detecting OOD samples with the output confidence. Later, some methods [23], [24], [25], [26] built the advanced detectors in a post-hoc manner. For instance, Liang et al. [23] combined temperature scaling and input preprocessing to achieve better detection performance. Instead of using the output confidence, Lee et al. [24] utilized the Mahalanobis distance between the test samples’ feature representations and the train samples’ . However, these methods require many labeled ID samples for training. There were some methods that focused on how to exploit unlabeled data for
found that self-supervised learning on the pure unlabeled ID data could improve the detection performance. For instance, Sehwag et al. [32] combined contrastive learning and Mahalanobis distance for OOD detection. There was also a line of works [33], [34], [35] which employed deep generative models on the pure unlabeled ID data. The mismatched samples in the unlabeled data can be filtered out OOD samples concurrently. Guo et al. [36] treated equally, hence the model still needed many labeled ID samples and filter out OOD samples concurrently. Yu et al. [37] considered the class distribution mismatch between labeled and unlabeled data. The mismatched samples in the unlabeled data can be regarded as OOD samples. Chen et al. [38] filtered out OOD samples in the unlabeled data with a confidence threshold and trained the model on the remaining data only. Yu et al. [39] tried to utilize mixed unlabeled data for OOD detection. It encouraged two classifiers to maximally disagree on the mixed unlabeled data. However, each unlabeled sample was treated equally, hence the model still needed many labeled samples to distinguish between ID and OOD samples. We have: 

\[
p(\hat{y} = C + 1|z) = 1 - p(y^*|z)
\]

where \(N\) is the nearest neighbors searching from \(Z^s\) and \(p(y = C + 1|z)\) means the supreme of \(p(y = C + 1|z)\). All samples in the feature space are normalized.

### 3.2 Empirical Estimation of Unknown Samples

To estimate the probability of a target sample \(z\) belonging to the unknown class, we first introduce an empirical estimation of the posterior probability \(p(y = C + 1|z)\).

**Theorem 1.** With the feature set of samples from two domain \(Z = Z^s \cup Z^t\) and a target features \(z \in Z^t\). If \(p_{C_{\text{ID}}}(z; k) = c_1\{\max_{i=0,...,C} \hat{p}_i(z; k, k_i | y = i) \leq \beta\} \) and \(k_i\) satisfies \(c_0 \frac{E(Z)}{E(Z)} \cdot k_i \leq \beta\) where \(k\) and \(k_i\) are the number of neighbors and the number of neighbors belonging to class \(i\), respectively. We have:

\[
\hat{p}(y = C + 1|z) = \max_{i=0,...,Y} \left( \eta \{y' = i\} \right)
\]

where \(N\) is the nearest neighbors searching from \(Z^s\) and \(p(y = C + 1|z)\) means the supreme of \(p(y = C + 1|z)\). All samples in the feature space are normalized.

**Proof 1.** We provide the proof sketch to show our key ideas which revolves around performing the empirical estimation of \(p(y = C + 1|z)\).

First, since we have no idea which known class is the source private class, we denote:

\[
p(y = C + 1|z) = 1 - p(y \in Y^s|z)
\]

we estimate the posterior distribution of \(z\) belonging to one of source classes which is easier.

By the Bayesian rule, the probability of \(z\) belonging to one of the source classes can be found as:

\[
p(y \in Y^s|z) = \frac{p(z|y \in Y^s) \cdot p(y \in Y^s)}{p(z)}
\]

\[
= \frac{\sum_{i=0}^{C} p_i(z) \cdot \sum_{j=0}^{C} p(y = j)}{\sum_{i=0}^{C} p_i(z) \cdot \sum_{j=0}^{C} p(y = j) + p_{C_{\text{ID}}}(z) \cdot p(y = C + 1)}
\]

Then, the estimation of the \(p(y \in Y^s|z)\) can boil down to deriving the estimations of probability density functions \(p_0(z), \ldots, p_{C_{\text{ID}}}(z)\).

Then, since \(z \in \mathbb{R}^n\) and all features are normalized where \(|z| = 1\), all data points locate on the surface of a \(m\)-dimensional unit sphere. We set \(U_r(z) = \{||z' - z||_2 \leq r\} \cap \{z' \in Z^s\}\), which is a set of data points from source domain on the unit hyper-sphere centered in the \(z\) with a radius \(r\). Assuming the density probability functions satisfy Lebesgue’s differentiation theorem, the probability density function can be attained by:

\[
p_i(z) = \lim_{r \to 0} \frac{p(z' \in U_{r_i}(z; r)|y = j)}{|U_{r_i}(z)|}
\]

Then, we donate \(r_k\) as the \(k\)-nearest neighbor distance which means the \(r_k\) Euclidean distance away from center \(z\) and its \(k - 1\)th nearest neighbor. Then we get:

\[
\hat{p}_i(z; k) = \frac{p(z' \in U_{r_k}(z)|z' \in Z^s)}{|U_{r_k}(z)|}
\]

With the denoting that \(B_i\) is the smallest sphere containing \(Z^s = \{z_1', \ldots, z_i'\}\), where \(Z^s\) is the set of all source sample belong to class \(i\), we assume that:

\[
\forall i, j \in \{0, \ldots, C\}; B_i \cap B_j = \emptyset
\]
Then, we have:

\[ \hat{p}(\mathbf{z}' \in U_{rk}(\mathbf{z})| \mathbf{z}' \in Z'_s) = \frac{|U_{rk}(\mathbf{z}) \cap B_i|}{|U_{rk}(\mathbf{z})|} \tag{7} \]

We assuming the number of neighbors are big enough. Then we have the estimation 

\[ |U_{rk}(\mathbf{z}) \cap B_i| = c_0 k_{i} \]  where 

\[ c_0 \] is a constant. Then the approximation of \( \hat{p}(\mathbf{z}; k) \) can be attained by:

\[ \hat{p}_i(\mathbf{z}; k, k_i) = c_0 \frac{k_i}{k(r_k)^{m-1}} \tag{8} \]

where \( k_i \) is the number of samples belonging to \( U_{rk} \) and \( B_i \).

Then, another challenge of estimating \( p(y \in Y^s|z) \) is computing \( p_{C+1} \) since we do not have a litter prior knowledge about unknown samples. The only knowledge we have is that samples not belonging to all the source classes belong to unknown class. Thus, we can attain:

\[ \hat{p}_{C+1}(\mathbf{z}; k) = c_1 \max_{i=0,\ldots,C} \hat{p}_i(\mathbf{z}; k, k_i) \leq \beta \tag{9} \]

where \( \beta \) is a constant chosen to satisfy the equation. According to Eqs.(2,3,8) and denoting \( \sum_{i=0}^{C} p(y = j) = \epsilon \) we have:

\[ p(y = C+1|z) = 1 - \frac{c_0}{\epsilon c_0 + (1 - \epsilon)c_1 k_{i}^{m-1}} \leq 1 - \frac{1}{\epsilon + (1 - \epsilon)c_1 k_i \epsilon/k_{i}} \tag{10} \]

where \( k_i = \max_{i=0,\ldots,C} k_i \). According to Eq.(10), we can finally get Eq.(1).

### 3.3 Unknown Detection via Linear Sub-space

By Theorem 1, the posterior probability of a target sample belonging to the unknown class depends on the biggest number of neighbors belonging to the same class. However, the misalignment between source domain and target domain can effect the accuracy of the estimation based on the neighborhood information. Some unknown samples can be close to one of the source classes while the labels of neighbors of a known sample can be unconscious due to the bias and less-discriminative embedding in two domains. Since it is hard to complete the domain misalignment without any prior knowledge about the unknown

Notably, in ODA problem where the source label set is contained in the target label set, we can get the equation:

\[ p(y \in Y^s|z) = p(y \in Y^{com}|z) \tag{11} \]

Then, based on Eqs.(3,8) we can estimate the probability of a target sample belonging to one of the common classes reliably but in OPDA, Eq.(11) would not work any more because of the private source classes existing.
samples and the private source classes in the original feature space. An unexpected category shift can lead to catastrophic misalignment where a whole class of target samples can be classified into the wrong source class [19]. Thus, except the domain misalignment, we propose to find a reliable linear sub-space of the original feature space which reduces the information redundancy of each sample due to the less-discriminative embedding to make them more sparse and adapt the misalignment between target samples and source samples. And finally it can improve the accuracy of the unknown detection based on the neighborhood information searched from source domain of a target sample in the sub-space. In specific, given the feature representations \( Z \in \mathbb{R}^{n \times m} \) where \( Z = M \cup B \) and \( B \) is the mini-batch of target samples, we have

\[
Z_{\text{sub}} = ZP, \quad \text{where} \quad P \in \mathbb{R}^{m \times p}, Z_{\text{sub}} \in \mathbb{R}^{n \times p}
\]  

where \( P \) is a transformation matrix mapping the features with \( m \) dimensions to reduced feature with \( p \) dimensions.

As illustrated in FIGG First, since the features transformation from a high dimension to a low dimension can lead to the information loss which might produce the new misalignment, we need to find the dimensions which represent the features of data points in the original feature space as much as possible. Therefore, we propose to maximize the variance of data points to find the major dimensions which can represent the main features of original data points and reduce the information loss after dimensionality reduction as much as possible to avoid generating the new misalignment between two domains. Then, to adapt the misalignment existing in the original feature space due to the less-discriminative embedding, we minimize the co-variance between \( z_i, z_j \in Z \) to decrease the correlation between two samples which can reduce the information redundancy of each samples to make the data points more sparse and discriminative. Then, we can figure out the minor dimensions which might cause the misalignment due to the less-discriminative embedding. Finally, we find the major dimensions which can represent the samples’ original features to the maximum extent and cut off the minor dimensions which are possible to bring the misalignment to get the reliable sub-space. Notably, we employ the Principal Component Analysis (PCA) to compute the transformation matrix.

To estimate the posterior distribution \( p(y = C + 1|z) \) by Theorem 1, we first employ a memory bank \( M \) to store all features of source samples:

\[
M^{(t)} = [m_1^{(t)}, \ldots, m_n^{(t)}]
\]  

we update the memory with a momentum scheme:

\[
m_i^{(t+1)} = \alpha m_i^{(t)} + (1 - \alpha)Z_i^{(t)}
\]  

By extracting crucial features from the original space, we propose to search neighbors of each target sample in the linear sub-space which has a more discriminative distribution of data points. Thus, with \( N \) for a target sample \( z \in \mathbb{R}^m \) we propose an estimation of \( p(y \in Y^{|z}) \) according to Eq. (1):

\[
\hat{p}(y \in Y^{|z}, k) = \max_{i=0, \ldots, Y} (\{z' \in N^k | y' = i\})
\]  

where \( k \) is the number of searched neighbors. Then, the detected unknown samples are donated as:

\[
\hat{Z}_\text{unk} = \{z | \hat{p}(y \in Y^{|z}) \leq \tau\}
\]  

where \( \tau \) is set manually.

**Discussions** According to Theorem 1, the posterior probability of a target sample belonging to the unknown class are also related to both the \( k \)-nearest neighbor distance. However, in the sub-space, the distribution of samples are much more sparse and uniform as shown in FIGG which is not discriminative enough to distinguish the unknown samples from known samples. Therefore, we select the more reliable scheme based on the biggest number of neighbors belonging to the same class which is robust against the influence the domain misalignment empirically. We also conduct experiments to compare the reliability of these two factors.

### 3.4 Learning

Since the classifier is trained to categorize a target sample into one of the source classes with highly confidence and distinguish target samples belonging to unknown class via the entropy of the outputs. Thus, the training of classifier \( D \) involves a trade-off as shown in Fig. 4 maximizing classification performance on \( Z^\text{com} \) and preventing overconfident predictions on \( Z^\text{unk} \). Since a traditional method [7], [13], [14] to improve the classification performance by training the classifier in the source domain and most of them employ a cross-entropy loss:

\[
\mathcal{L}_S = \frac{1}{n} \sum_{z_i \in Z} - \log e^{W_{z_i}z_i} + \sum_{j \neq i} e^{W_{z_i}z_i}
\]  

But this method can usually lead to the overconfident predictions on unknown samples [9]. To deal with the problem, Cao et al. [7] aggregated multiple complementary uncertainty measures, OVANet [13] employ an one-vs-all classifier for unknown detection and GATE [9] proposed an energy-based classifier by extending the traditional softmax-based classifier to improve the classification performance on distinguishing the unknown samples. In this work, without applying extra parameters, we propose a more efficient method to train only one classifier via three losses.
Algorithm 1: Exploiting Inter-Sample

**Requirement:**
Source dataset \(Z^s, Y^s\), target dataset \(Z^t\),
The number of neighbors \(k\) and the threshold \(\tau\)

**Training:**
while step < max step do
    if step = 0 do
        Extract all features from \(Z^s\) and build the memory \(M\)
        Sample batch \(B^s\) from \((Z^s, Y^s)\) and batch \(B^t\) from \(Z^t\)
        Extract features from each of \(B^s\) and \(B^t\)
    for \(z_i \in B^s\) do
        Update the memory by Eq. 14
    for \(z_i \in B^t\) do
        Retrieve the nearest neighbors \(N_i\) for \(z_i\) from \(M\)
        Compute \(\hat{p}(y \in Y^s | z_i, k)\)
        if \(\hat{p}(y \in Y^s | z_i, k) < \tau\) do
            Append \(z_i\) into \(Z^{unk}\)
        Compute \(L_{sup}\) based on Eq. 21
        Compute the margin \(\mu\) based on Eq. 19
        Compute \(L_{margin}\) based on Eq. 18
        Compute \(L_t\) based on Eq. 20
        Compute the overall loss \(L_{all}\)
    Update the model

3.4.1 Unknown-adaptive margin loss (UAM)

Since using the traditional CE loss training on source domain creates an imbalance problem between the known and unknown classes for target samples. Learning on source classes \(Y^s\) is much faster than that on the unknown class \(Y^{unk}\) due to the supervised objective which leads to the whole model to be biased towards the common classes \(Y^s\) in target domain. Then, the predictions of many unknown samples can be overconfident. Therefore, it is important to slow down the convergence rate of the predictions. We propose a new unknown-adaptive margin loss (UAM) to control the gradient descent rate of source classes based on the confidence level of detected unknown target samples in the current mini-batch:

\[
L_{uam} = \frac{1}{n} \sum_{z_i \in Z^t} - \log \frac{e^{W^T_{y^s \cdot z_i + \alpha \mu}}}{\sum_{j \neq i} e^{W^T_{y^s \cdot z_i}}} \tag{18}
\]

where \(\mu\) is computed by the difference value between the entropy of predictions of target samples in \(Z^{unk}\) and the entropy threshold \(\delta\). With donating \(S(C(z_i))\) as the entropy of the prediction of \(z_i\):

\[
\mu = \frac{1}{|Z^{unk}|} \sum_{z_i \in Z^{unk}} \max(0, \delta - S(C(z_i))) \tag{19}
\]

where \(|\cdot|\) means the number of elements in a set and the \(\delta\) is the threshold used in the test phase to determine whether a target sample is unknown sample. Since \(\log C\) is the maximum of \(E(C(z_i))\), we set \(\delta = \frac{\log C}{2}\). \(\mu\) can represent the difference between the confidence level of detected unknown samples and the threshold \(\delta\) at current step. If \(\mu\) is big, the predictions of the unknown samples are overconfident which means we need to slow down the learning on the source domain the learning on source domain.

3.4.2 Unknown loss

Since we use the entropy of the classifier output to distinguish the unknown samples, to lower the confidence of unknown samples belonging to \(\hat{Z}^{unk}\), we employ the Kullback-Leibler divergence to a uniform class posterior distribution for unknown inputs to increase the entropy:

\[
L_{unk}(z) = -\frac{1}{2} \sum_{i=0}^{Y} \log D(z) - \log Y \tag{20}
\]

where the inputs of Eq. (20) belong to the detected unknown samples set \(\hat{Z}^{unk}\).

3.4.3 Supervised contrastive loss on source domain

Moreover, to reduce the influence of the domain misalignment, it is necessary to make each source class more compact and discriminative to enlarge the gap between two source classes which can improve the consistency of neighbors' labels for a target sample when it searches the nearest neighbors from source domain. Thus, we employ a supervised contrastive learning loss \(L_{sup}\) to make data points from source domain more compact by pushing the samples from different classes away while pulling the samples from the same class closer. \(L_{sup}\) is given by:

\[
L_{sup} = \frac{1}{|B^s|} \sum_{i=0}^{Y} \sum_{m_j \in A^+_i \cup A^-_i} \exp(<z_i, m_j>/t) \tag{21}
\]

where \(A^+_i\) and \(A^-_i\) represent the positive samples in \(M\) with the same label as \(z_i\) and the negative samples searched from \(M\). \(t\) is a temperature parameter.

3.4.4 Total loss and algorithm

The total training loss of our method can be computed as

\[
L_{all} = L_{uam} + L_{sup} + \lambda L_{unk} \tag{22}
\]

where \(\lambda\) is a weighting parameter. Moreover, the full algorithm of our method is provided in Algorithm 1.

4 Experimental Results

In this section, we first introduce our experimental setups, including datasets, evaluation protocols and training details. Then, we compare our method with the SOTA UDA methods. We also conduct extensive ablation studies to demonstrate the effectiveness of each component of the proposed method. All experiments were implemented on one RTX2080Ti 11GB GPU with PyTorch 1.7.1.

| Method | Office31 (10/10/11) |
|--------|----------------------|
| UAN [17] | 59.7 58.6 60.1 70.6 71.4 60.3 63.5 |
| CMU [15] | 68.1 67.3 71.4 79.3 80.4 72.2 73.1 |
| DANCE [19] | 78.6 71.5 79.9 91.4 87.9 72.2 80.3 |
| DCC [14] | **88.5** 78.5 70.2 79.3 88.6 75.9 80.2 |
| ROS [16] | 71.4 71.3 81.0 94.6 95.3 79.2 82.1 |
| USFDA [41] | 85.5 79.8 83.2 90.6 88.7 81.2 84.8 |
| OVNNet [13] | 85.8 79.4 80.1 95.4 94.3 84.0 86.5 |
| GATE [18] | 87.7 81.6 **84.1** 94.8 91.4 83.4 87.6 |

Ours 87.3 83.5 82.2 96.1 **99.2** 84.7 **88.8**

**TABLE 1**

Results on Office31 with OPDA setting (H-score)
### 4.1 Experimental setups

#### 4.1.1 Datasets and evaluation protocols.

We conduct experiments on four datasets. Office-31 [43] consists of 4,652 images from three domains: DSLR (D), Amazon (A), and Webcam (W). OfficeHome [44] is a more challenging dataset, which consists of 15,500 images from 65 categories. It is made up of 4 domains: Artistic images (Ar), Clip-Art images (Cl), Product images (Pr), and Real-World images (Rw). VisDA [45] is a large-scale dataset, where the source domain contains 15,000 synthetic images and the target domain consists of 5,000 images from the real world.

In this paper, we use the H-score in line with recent UDA methods [13, 14, 15]. H-score, proposed by Fu et al. [15], is the harmonic mean of the accuracy on the common classes $a_{com}$ and the accuracy on the unknown class $a_{unk}$:

$$h = \frac{2a_{com} \cdot a_{unk}}{a_{com} + a_{unk}}. \quad (23)$$

#### 4.1.2 Training details.

We employ the ResNet-50 [46] backbone pretrained on ImageNet [47] and optimize the model using Nesterov momentum SGD with momentum of 0.9 and weight decay of $5 \times 10^{-4}$. The batch size is set to 36 through all datasets. The initial learning rate is set as 0.1 for the new layers and 0.001 for the backbone layers. The learning rate is decayed with the inverse learning rate decay scheduling. The number of neighbors retrieved is set to be dependent on the sizes of the dataset. For Office-31 (4,652 images in 31 categories) and OfficeHome (15,500 images in 65 categories), the number of retrieved neighbors $|\mathcal{N}_i|$ is set to 10 and $\tau$ is set to 6. For VisDA (20,000 images in total), we set $|\mathcal{N}_i|$ to 100 and $\tau$ is set to 6, respectively. We set $\lambda$ to 0.1 and $t$ to 0.05 which are constant through all the datasets. In the test phase, the threshold of distinguishing the unknown samples is set to $\log C - 2$ following DANCE [19] which used the entropy of the classifier output to determine the unknown samples.

### 4.2 Comparison with the SOTA Methods

#### 4.2.1 Baselines.

We aim to show that our method can better balance the confidences of known and unknown samples for UniDA by comparing our method with the current SOTA methods, such as UAN [17] and DANCE [19], which employed a softmax-based classifier to produce the confidence of each sample to determine whether it belongs to the unknown class or not. Also, we compare our method with DANCE [19], DCC [13] and GATE [18] to show that it is more effective to adapt the domain misalignment through the dimensionality reduction rather than completing it in the original feature space.

#### 4.2.2 Results in main datasets.

Tables 1 and 2 list the results on Office-31 with OPDA setting and ODA setting respectively. Tables 3 and 4 list the results on OfficeHome and VisDA both with OPDA setting and ODA setting, respectively. On Office-31, our method outperforms the SOTA methods by 1.2% in terms of the H-score on average with OPDA setting and made a significant improvement of 2.0% in terms of the H-score on average with ODA setting. For the more challenging dataset OfficeHome which contains much more private classes than common classes, our method also made a significant improvement of 1.0% in terms of the H-score. Our method consistently performs better than other methods. VisDA is a much larger dataset than Office-31 and OfficeHome which contains about 10000 images in each domain.

#### 4.2.3 Summary.

According to the results of quantitative comparisons, our method achieves the SOTA performance in every dataset and most sub-tasks, which demonstrates the main idea of our method that adapt the domain misalignment through dimensionality reduction and balance the confidence of known and unknown samples by controlling the gradient updating of learning on the source domain.

### 4.3 Ablation Studies

#### 4.3.1 A closer look at the unknown detection

How to choose the confidence for unknown detection

By Theorem 1, the posterior probability of a target sample belonging to the unknown class depends on the biggest number of neighbors belonging to the same class. Thus, we compare the proposed unknown detection scheme with the detection based on the $k$-nearest neighbor distance in this part. First, we collect the $k$-nearest neighbor distances of all target samples in an early epoch in Office-31 which is more influential for the whole training. As shown in the first row in Fig. 5, the distributions of known samples are not distinguishable enough especially in A2W, D2A and W2A. It is obvious that the $k$-nearest neighbor distance is not reliable enough to detect the unknown samples. Moreover, the optimal thresholds for each sub-tasks are different and hard to choose. Notably, mapping samples into the linear sub-space has the significant influence in uniforming the distribution of data points and make them more sparse as shown in the second row in Fig. 5. However, it is helpless to improve the discrimination of unknown samples based on the $k$-nearest neighbor distance. By contrast, the distributions of confidences defined by the biggest number of neighbors belonging to the same class illustrated in the first row of Fig. 5 are much more distinguishable.
4.3.2 Effect of sub-space

Original feature space VS Linear Sub-space  The purpose of extracting a linear sub-space is to make the distribution of data points more sparse and uniform to avoid the influence of the domain misalignment. Comparing the charts in the first row and the second row in Fig. [5], it is obviously that the data points in the sub-space are similar in the distribution of the $k$-nearest neighbors. In Fig. [8], we conduct experiments on Office-31 to compare the distributions of the confidences produced by the proposed unknown detection method in the original feature space (First row) and the linear sub-space (Second row). Apparently, the distributions of confidences in the sub-space are much more distinguishable than that in the original feature space where the overlaps between unknown and known samples are much fewer in the sub-space which is benefited by reducing the misalignment between target samples and source samples.

Accuracy of the proposed unknown detection  We also conduct experiments on the accuracy of the unknown detection on the original feature space and the sub-spaces with different dimensions on some sub-tasks of Office-31 and OfficeHome where we plot the histograms of the accuracy of the unknown detection in Fig. [6]. The accuracy of unknown samples detection is consistently detected with high accuracy which far surpasses 80% on average and the performance of accuracy of unknown detection in the sub-space is much better than that in the original feature space. Thus, through the proposed unknown detection scheme, our approach finds the unknown samples in the target domain reliably.
10

Fig. 6. Histograms of accuracy of unknown detection in the original feature space (red columns) and in the sub-space (green columns) on Office-31 (A2D, A2W, D2A, D2W and W2A) and OfficeHome (A2C, A2P, C2A, C2R and P2R) with the OPDA setting.

Fig. 7. The t-SNE visualizing cooperation. Different colors represent different classes. Red points represent the unknown samples while the points in other colors represent the known samples of different classes.

| Method         | Office-31 (10/10/11) | A2D | A2W | D2A | D2W | W2D | W2A | Avg |
|----------------|-----------------------|-----|-----|-----|-----|-----|-----|-----|
| w/o $\mathcal{L}_{sup}$ | 81.0 77.5 78.2 | 95.0 | 91.0 | 72.9 | 82.6 |
| w/o $\mathcal{L}_{uam}$  | 79.0 81.0 78.1 | 94.4 | 99.7 | 74.1 | 84.4 |
| w/o $\mathcal{L}_{unk}$ | **89.5** | 84.9 | 89.7 | 93.7 | 85.8 | 88.5 | 88.7 |

**TABLE 5**
Effect on performance of each loss in Office with the OPDA setting.

| Method | Office-31 (10/10/11) | A2D | A2W | D2A | D2W | W2D | W2A | Avg |
|--------|-----------------------|-----|-----|-----|-----|-----|-----|-----|
| OSBP [11] | 81.0 77.5 78.2 | 95.0 | 91.0 | 72.9 | 82.6 |
| ROS [16] | 79.0 81.0 78.1 | 94.4 | 99.7 | 74.1 | 84.4 |
| Ovanet [13] | **89.5** | 84.9 | 89.7 | 93.7 | 85.8 | 88.5 | 88.7 |
| Ours | **89.9** | 83.6 | **92.0** | 94.1 | 95.5 | **91.2** | **91.0** |

**TABLE 6**
Results on Office-31 using the VGGNet [48] backbone with the ODA setting.

4.3.3 **Effect of losses**

We conduct ablation experiments to validate the effect of each loss. As shown in Table 5 we first list the effect on the performance of Office with OPDA setting. We can observe that each loss

**Unknown-adaptive margin loss VS CE-loss** To show the effect of the unknown-adaptive margin loss on balancing the predictions of known and unknown samples, we track the entropy level of unknown samples and the confidence level of known samples following the training process on Office-31. We recorded the mean value of entropies of predictions for unknown samples and the mean value of maximum prediction confidences output by the classifier after softmax of each known samples in every steps. For comparison, we first plot records where the classifier was only trained on source domain with a traditional CE-lass like Eq. 17 in the first row of Fig. 9. We can observe that although the known samples have high confidence consistently, the unknown samples are significantly overconfident during the training. In the second row of Fig. 8, we plot the records using the CE-loss and proposed KL divergence loss as Eq. 20. The entropy level has been improved but the overconfidence of unknown samples is still serious.

In the third row of Fig. 9, we plot the records using the proposed unknown-adaptive loss which can perfectly distinguish the unknown samples with entropy while keeping the known samples with highly confident he predictions. Notably, to better show the effect of the proposed losses, we only compute the mean entropy of the detected unknown samples updated by the unknown loss in the second row charts. We also employ t-SNE pictures to visualize the distributions of target samples in Fig. 7. We can observe that the distribution of data points in ours (right) is much more discriminative than that of DANCE [19](mid) and the model training with the source dataset only (left).
Setting of unknown-adaptive margin. To show the effect of the unknown-adaptive margin selection scheme, we compare it with the human-picked thresholds on Office-31(A2D, D2A) and OfficeHome(R2C). From Fig. 10(a), we can observe that it is difficult to choose a consistently optimal threshold for all datasets and sub-tasks as the model is sensitive to the thresholds.

4.3.4 Performance on VGGNet [48].

Table 3 shows the quantitative comparison with the ODA setting on Office-31 using VGGNet [48] instead of ResNet50 as the backbone for feature extraction. According to the results, we demonstrate that our method is also effective with another backbone without changing any hyper-parameters.

4.3.5 Sensitivity to hyper-parameters

To show the sensitivity of $\lambda$ in the total loss and the human set threshold $\tau$, we conducted experiments on Office-31 with the OPDA setting. Fig. 10(a) shows that the change of $\tau$ has little influence on the performance when $\tau > 0.4$. Fig. 10(b) shows that our method has a highly stable performance over different values of $\lambda$. 
5 Conclusion

In this paper, we propose a new framework to reduce the influence of the domain misalignment and balance the predictions of known and unknown target samples. Its core idea is to estimate the probabilities of target samples belonging to the unknown class by the biggest number of neighbors with the same label searched from the source domain and detect the unknown samples via mapping the features in the original feature space into a linear sub-space to make the data points more sparse and discriminative which can reduce the influence of domain misalignment. Also, our method well balance the confidences of known samples. Thus in the future work, we plan to extend our method to leverage this relationship for further boosting the performance with the OPDA and ODA setting.

A limitation of our method is that it does not sufficiently utilize the inter-sample relationship within the set of unknown samples. Thus in the future work, we plan to extend our method to leverage this relationship for further boosting the performance with the OPDA and ODA setting.

References

[1] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in International conference on machine learning. PMLR, 2015, pp. 1180–1189.

[2] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, “Maximum classifier discrepancy for unsupervised domain adaptation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3723–3732.

[3] M. Long, H. Zhu, J. Wang, and M. I. Jordan, “Unsupervised domain adaptation with residual transfer networks,” arXiv preprint arXiv:1602.04433, 2016.

[4] B. Gong, K. Grauman, and F. Sha, “Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation,” in International Conference on Machine Learning. PMLR, 2013, pp. 222–230.

[5] Y. Zou, Z. Yu, B. Kumar, and J. Wang, “Unsupervised domain adaptation for semantic segmentation via class-balanced self-training,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 289–305.

[6] Z. Cao, L. Ma, M. Long, and J. Wang, “Partial adversarial domain adaptation,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 135–150.

[7] Z. Cao, K. You, M. Long, J. Wang, and Q. Yang, “Learning to transfer examples for partial domain adaptation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2985–2994.

[8] J. Zhang, Z. Ding, W. Li, and P. Ogunbona, “Importance weighted adversarial nets for partial domain adaptation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8156–8164.

[9] J. Liang, Y. Wang, D. Hu, R. He, and J. Feng, “A balanced and uncertainty-aware approach for partial domain adaptation,” in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. Springer, 2020, pp. 123–140.

[10] P. Panareda Busto and J. Gall, “Open set domain adaptation,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 754–763.

[11] K. Saito, S. Yamamoto, Y. Ushiku, and T. Harada, “Open set domain adaptation by backpropagation,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 153–168.

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

[13] K. Saito and K. Saenko, “Ovanet: One-vs-all network for universal domain adaptation,” arXiv preprint arXiv:2104.03344, 2021.

[14] G. Li, K. Gang, Y. Zhu, Y. Wei, and Y. Yang, “Domain consensus clustering for universal domain adaptation,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[15] B. Fu, Z. Cao, M. Long, and J. Wang, “Learning to detect open classes for universal domain adaptation,” in European Conference on Computer Vision. Springer, 2020, pp. 567–583.

[16] S. Bucci, M. R. Loughman, and T. Tommasi, “On the effectiveness of image rotation for open set domain adaptation,” in European Conference on Computer Vision. Springer, 2020, pp. 422–438.

[17] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan, “Universal domain adaptation,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 2720–2729.

[18] L. Chen, Y. Lou, J. He, T. Bai, and M. Deng, “Geometric anchor correspondence mining with uncertainty modeling for universal domain adaptation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16 134–16 143.

[19] K. Saito, D. Kim, S. Sclaroff, and K. Saenko, “Universal domain adaptation through self supervision,” arXiv preprint arXiv:2002.07953, 2020.

[20] Z. Cao, M. Long, J. Wang, and M. I. Jordan, “Partial transfer learning with selective adversarial networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2724–2732.

[21] Q. Feng, G. Kang, H. Fan, and Y. Yang, “Attract or distract: Exploit the margin of open set,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7990–7999.

[22] D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” arXiv preprint arXiv:1610.02136, 2016.

[23] S. Liang, Y. Li, and R. Srikant, “Enhancing the reliability of out-of-distribution image detection in neural networks,” arXiv preprint arXiv:1706.02690, 2017.

[24] K. Lee, K. Lee, H. Lee, and J. Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” Advances in neural information processing systems, vol. 31, 2018.

[25] C. S. Sastry and S. Oore, “Detecting out-of-distribution examples with gram matrices,” in International Conference on Machine Learning. PMLR, 2020, pp. 8491–8501.

[26] Y. C. Hsu, Y. Shen, H. Jin, and Z. Kira, “Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10 951–10 960.

[27] D. Hendrycks, M. Mazeika, and T. Dietterich, “Deep anomaly detection with outlier exposure,” arXiv preprint arXiv:1812.04606, 2018.

[28] I. Golan and R. El-Yaniv, “Deep anomaly detection using geometric transformations,” Advances in neural information processing systems, vol. 31, 2018.

[29] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, “Using self-supervised learning can improve model robustness and uncertainty,” Advances in neural information processing systems, vol. 32, 2019.

[30] J. Winkens, R. Bunel, A. G. Roy, R. Stanforth, V. Natarajan, R. J. Ledsam, P. MacWilliams, P. Kohli, A. Karthikesalingam, S. Kohl et al., “Contrastive training for improved out-of-distribution detection,” arXiv preprint arXiv:2007.05566, 2020.

[31] J. Tack, S. Mo, J. Jeong, and J. Shin, “Csi: Novelty detection via contrastive learning on distributionally shifted instances,” Advances in neural information processing systems, vol. 33, pp. 11 839–11 852, 2020.

[32] V. Sehwag, M. Chiang, and P. Mittal, “Ssd: A unified framework for self-supervised outlier detection,” arXiv preprint arXiv:2103.12051, 2021.

[33] E. T. Nalisnick, A. Matsukawa, Y. W. Teh, and B. Lakshminarayanan, “Detecting out-of-distribution inputs to deep generative models using a test for typicality.” 2019.

[34] Y. Huang, S. Dai, T. Nguyen, R. G. Baraniuk, and A. Anandkumar, “Out-of-distribution detection using neural rendering generative models,” arXiv preprint arXiv:1907.04572, 2019.

[35] J. Serrà, D. Álvarez, V. Gómez, O. Slizovskaia, J. F. Núñez, and J. Luque, “Input complexity and out-of-distribution detection with likelihood-based generative models,” arXiv preprint arXiv:1909.11480, 2019.

[36] Y. Huang, S. Dai, T. Nguyen, R. G. Baraniuk, and A. Anandkumar, “Out-of-distribution detection using neural rendering generative models,” arXiv preprint arXiv:1909.11480, 2019.
[37] Q. Yu, D. Ikami, G. Irie, and K. Aizawa, “Multi-task curriculum framework for open-set semi-supervised learning,” in European Conference on Computer Vision. Springer, 2020, pp. 438–454.

[38] L.-Z. Guo, Z.-Y. Zhang, Y. Jiang, Y.-F. Li, and Z.-H. Zhou, “Safe deep semi-supervised learning for unseen-class unlabeled data,” in International Conference on Machine Learning. PMLR, 2020, pp. 3897–3906.

[39] Q. Yu and K. Aizawa, “Unsupervised out-of-distribution detection by maximum classifier discrepancy,” in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9518–9526.

[40] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” arXiv preprint arXiv:1710.09412, 2017.

[41] J. N. Kundu, N. Venkat, R. V. Babu et al., “Universal source-free domain adaptation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4544–4553.

[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., “Pytorch: An imperative style, high-performance deep learning library,” Advances in neural information processing systems, vol. 32, pp. 8026–8037, 2019.

[43] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, “Adapting visual category models to new domains,” in European conference on computer vision. Springer, 2010, pp. 213–226.

[44] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, “Moment matching for multi-source domain adaptation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1406–1415.

[45] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko, “Visda: The visual domain adaptation challenge,” arXiv preprint arXiv:1710.06934, 2017.

[46] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[48] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[49] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.” Journal of machine learning research, vol. 9, no. 11, 2008.