

Unknown-Oriented Learning for Open Set Domain Adaptation

Jie Liu^[0000-0002-1327-1315], Xiaoqing Guo^[0000-0002-9476-521X], and
Yixuan Yuan [✉][0000-0002-0853-6948]*

City University of Hong Kong
{jliu.ee, xqguo.ee}@my.cityu.edu.hk yxyuan.ee@cityu.edu.hk

Abstract. Open set domain adaptation (OSDA) aims to tackle the distribution shift of partially shared categories between the source and target domains, meanwhile identifying target samples non-appeared in source domain. The key issue behind this problem is to classify these various unseen samples as unknown category with the absent of relevant knowledge from the source domain. Though impressing performance, existing works neglect the complex semantic information and huge intra-category variation of unknown category, incapable of representing the complicated distribution. To overcome this, we propose a novel Unknown-Oriented Learning (UOL) framework for OSDA, and it is composed of three stages: true unknown excavation, false unknown suppression and known alignment. Specifically, to excavate the diverse semantic information in unknown category, the multi-unknown detector (MUD) equipped with weight discrepancy constraint is proposed in true unknown excavation. During false unknown suppression, Source-to-Target grAdient Graph (S2TAG) is constructed to select reliable target samples with the proposed super confidence criteria. Then, Target-to-Target grAdient Graph (T2TAG) exploits the geometric structure in gradient manifold to obtain confident pseudo labels for target data. At the last stage, known alignment, the known samples in the target domain are aligned with the source domain to alleviate the domain gap. Extensive experiments demonstrate the superiority of our method compared with state-of-the-art methods on three benchmarks.

Keywords: Domain Adaptation, Open Set, Graph

1 Introduction

Deep learning has made spectacular progress in diverse application such as face recognition [24], medical image understanding [3, 11] and autonomous driving [46, 4]. These models are commonly trained under the supervised learning with

* ✉Yixuan Yuan is the corresponding author

This work was supported by Hong Kong Research Grants Council (RGC) General Research Fund 11211221(CityU 9043152), and Hong Kong Research Grants Council (RGC) Early Career Scheme grant 21207420 (CityU 9048179).

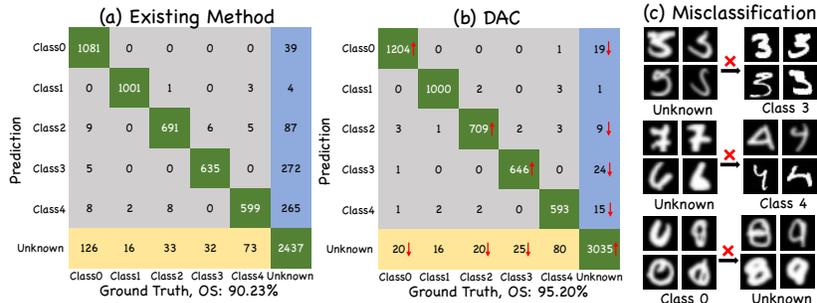


Fig. 1: The digital recognition confusion matrix of (a) existing method [23] and (b) our method. (c) Misclassification Example.

assumption that the training and test data come from the same distribution. However, this assumption, in practice, can be easily violated due to the change of environment or data acquisition device, leading to degraded performance in other domains. Unsupervised Domain Adaptation (UDA) techniques are proposed to alleviate this degeneration problem via narrowing down the domain gap. The main paradigm in UDA minimizes discrepancy metrics to reduce the distribution divergence [27, 22, 16], and the other paradigm leverages the adversarial learning to learn domain-invariant feature representations [20, 45, 21].

The aforementioned UDA algorithms work competently under the closed set regime, a.k.a. closed set domain adaptation (CSDA), where the label space shares in two domains ($\mathcal{C}_s = \mathcal{C}_t$). However, the target domain always contains unknown category (unseen in the source domain) in real-world scenarios, severely restricting the applicability of CSDA models. For example, a medical diagnosis intelligent system should recognize data as unknown when data belonging to an unseen category is accidentally inputted. The corresponding decision process should be transferred to the surgeon for safety diagnosis. Therefore, a more realistic setting, Open Set Domain Adaptation (OSDA) [34, 37], is recently studied, wherein the target domain contains irrelevant categories not presented in the source domain ($\mathcal{C}_s \subset \mathcal{C}_t$) [37]. Compared with CSDA, the goal of OSDA is not only to adapt the model from source to target domain for precise predictions within known categories, but also to reject unseen samples as unknown category.

To distinguish unknown samples from known ones, existing OSDA methods employ an additional logit to indicate the unknown [34, 37, 23, 1, 8]. However, the unknown category, containing a set of unseen classes, possesses complex visual semantics and huge intra-category variation. For example, digits 0 to 4 are known categories and 5 to 9 are unseen classes in digital recognition task. This phenomenon make the single classifier weight incapable of representing the dispersed distribution of unknown category, leading to misclassification of unknown samples into known categories and high false positive rate for known categories, as shown in the blue region of Fig. 1 (a). In addition, since there are abundant and complex semantics in the unknown category, some known samples

with similar semantics are classified into unknown category indistinguishably (yellow region), e.g., some digits 0 with similar semantic information to digit 8 are misclassified into unknown, as shown in Fig. 1 (c). These two types of overwhelming false predicted data would significantly hinder the practicability of machine learning system. Thus, unknown learning is a crucial problem in OSDA, but it has seldom been specially investigated yet. Here, we post a question: ‘How could we discover as many unknown samples as possible without misclassifying known samples into unknown category?’

Along with such a question, we propose a novel unknown-oriented learning (**UOL**) framework to solve OSDA, including *true unknown excavation*, *false unknown suppression* and *known alignment* three stages. To handle the complex visual semantics and huge intra-category variation, we propose the Multi-Unknown Detector (MUD) with a weight discrepancy constraint to detect the unseen samples scattered around the feature space and represent the dispersed distribution of unknown category in *true unknown excavation*. The weight discrepancy constraint enables the MUD to capture diverse semantic information of unknown category and excavate as many unknown samples as possible, which reduces the false positive rate for known category. Considering that the gradient of model parameters contains recognition-relevant information [30] and directional information [18], we advance two gradient graphs named Source-to-Target gradient Graph (S2TAG) and Target-to-Target gradient Graph (T2TAG) in *false unknown suppression* to fully delve into the knowledge of known samples in gradient space. Through exploiting the intrinsic relationship among samples in gradient space, we rectify the false unknown samples in the first stage and enhance the performance of known categories. Specifically, the S2TAG bridges the source domain with target domain, and selects confident target samples with the proposed super confidence criteria. Then, the T2TAG is devised to propagate the confidence of these target samples in the gradient manifold and obtain pseudo labels through solving a defined energy function. After distinguishing unknown samples, the scenario of OSDA is simplified as the CSDA problem. The *known alignment* stage is proceeded to alleviate the domain gap between the shared categories. In summary, our major contributions are summarized as follows:

- To explicitly learn the unknown category, a novel framework **UOL** is designed to conquer the OSDA problem with three stages, *true unknown excavation*, *false unknown suppression* and *known alignment*.
- In *true unknown excavation*, multi-unknown detector equipped with weight discrepancy constraint is proposed to explore the diverse semantic information of unknown. In *false unknown suppression*, two gradient graphs are constructed to obtain confident pseudo labels by exploiting the essential geometric structure of data.
- Extensive experiments are carried out on two standard OSDA benchmarks and one newly built medical diagnosis benchmark. The results demonstrate the superiority compared with other SOTA methods. Moreover, ablation studies validate the impact and effectiveness of the proposed **UOL** framework.

2 Related Work

2.1 Closed Set Domain Adaptation (CSDA)

CSDA is intended to alleviate the performance degradation caused by domain discrepancy. Existing methods could be mainly categorized into two streams: adversarial learning based methods [5, 45, 10, 41] and discrepancy metric based methods [38, 22, 16]. The first stream trains a domain discriminator to distinguish the source and target domains, and encourages feature extractor to learn domain-invariant feature via fooling the discriminator. Ganin et al. [10] advance a pioneering work, where a gradient reversal layer is proposed to connect the feature extractor and domain classifier. Subsequently, many adversarial based methods [39, 36, 42, 41, 5] are well designed. The second stream explicitly measures the domain gap by discrepancy metrics and narrows down the domain gap via minimizing the metrics [26, 44, 22, 38]. For instance, Long et al. [25] minimize the multi-kernel maximum mean discrepancy between two domains. Zhu et al. [27] propose the joint maximum mean discrepancy among multiple layers. While in real-world applications, open set is a more common scenario. But these CSDA methods fail in solving the problem of open set domain adaptation since they are incapable of rejecting unknown samples and the alignment of shared known distributions will be destructed.

2.2 Open Set Domain Adaptation (OSDA)

The purpose of open set domain adaptation is to reject unknown samples in target domain and align distributions of known categories between two domains, with the absent target annotations. Saito et al. [37] raise the realistic OSDA setting where unknown samples are only existing in the target domain. They train a classifier to build a boundary between source and target samples and train a generator to make target samples far from the boundary, so that unknown samples would be rejected and known distributions are aligned. In addition, Liu et al. [23] adopt a coarse-to-fine weighting mechanism to separate the target into known and unknown. Recently, subsequent works [8, 33, 43, 28, 1, 14, 35] introduce various techniques to generate more discriminative features in the target domain and benefit the OSDA problem. Feng et al. [8] emphasize the semantic structure of open set data via contrastive mapping. Pan et al. [33] employ the self-ensemble framework with category-agnostic clustering in the target domain and exploit the underlying structure of each cluster to learn more discriminative target information. Xu et al. [43] couple all data optimally and propose a prototypical loss to achieve intra-class compactness and inter-class separability. Luo et al. [28] utilize graph convolution to propagate the category information and achieve a tighter upper bound of the target error. Bucci et al. [1] use the inherent properties of self-supervision to achieve a more robustness recognition. Jing et al. [14] project the feature to a hyper-spherical latent space and constrain the centroid deviation angle to align the distribution.

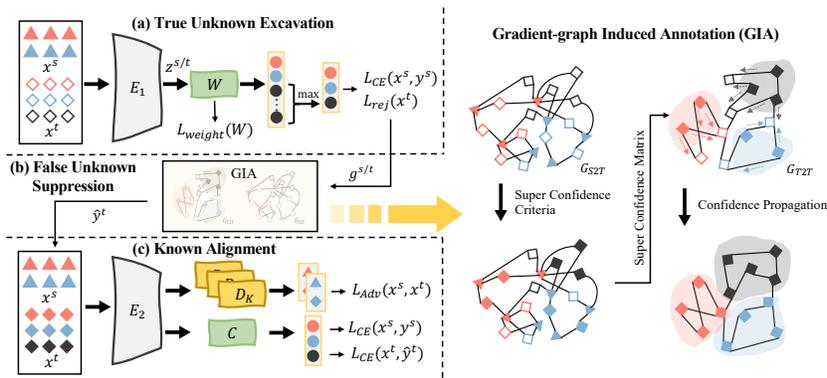


Fig. 2: Overview of the proposed framework, UOL, which includes three stages, i.e. (a) *True Unknown Excavation*, (b) *False Unknown Suppression* and (c) *Known Alignment*. In (a), source data \mathbf{x}^s and target data \mathbf{x}^t are fed into backbone E_1 to extract feature \mathbf{z} . Then, the multi-unknown detector W distinguishes the samples from diverse unseen categories and initially aligns the distribution of known categories. In (b), the gradients $\hat{\mathbf{g}}^{s/t}$ derived from back-propagation are utilized to generate pseudo labels \hat{y}^t through GIA, where S2TAG G_{S2T} is built to produce super confidence matrix and T2TAG G_{T2T} is built to propagate confidence in target domain. In (c), category-level discriminators $\{D_i\}_{i=1}^K$ align the feature distributions in category level, and extended classifier recognizes data into $K + 1$ categories.

However, the complicated unknown category distribution makes it hard to reject the unseen samples completely, which negatively influences the distribution alignment of known categories. Differently, we introduce multi-unknown detector with the help of discrepancy constraint to explore the diverse semantic information for unknown category. In addition, two graph structures are advanced to rectify the false prediction, so as to avoid the deconstruction from unknown samples.

3 Unknown-Oriented Learning

In OSDA, we have a source domain $\mathcal{D}^s = (\mathbf{x}_i^s, y_i^s)_{i=1}^{n_s}$ of n_s labeled samples associated with \mathcal{C}_s label space and a target domain $\mathcal{D}^t = (\mathbf{x}_i^t)_{i=1}^{n_t}$ of n_t unlabeled target samples associated with \mathcal{C}_t label space. The source label space \mathcal{C}_s is a subset of target label space \mathcal{C}_t , i.e. $\mathcal{C}_s \subset \mathcal{C}_t$, while target label space further includes a set of additional categories \mathcal{C}_{unk} , i.e. $\mathcal{C}_t = \mathcal{C}_s \cup \mathcal{C}_{unk}$. These unseen categories should be recognized as ‘unknown’ category in OSDA. Assuming the source and target domains are drawn from the distributions p and q respectively, where $p \neq q$, we further observe that probability distributions of shared categories are different, i.e. $p \neq q_{\mathcal{C}_s}$, due to the domain gap. In summary, we face

a label space shift problem $\mathcal{C}_s \neq \mathcal{C}_t$ and a distribution shift problem $p \neq q_{\mathcal{C}_s}$ in OSDA. Thus, this study aims to reject unseen samples as unknown category for the label space shift problem and to alleviate the domain gap of remaining samples for the distribution shift problem.

To ensure the distribution alignment among shared categories instead of mixing with unknown categories, we reject unseen data first (*true unknown excavation* and *false unknown suppression*) and then align distributions of the remaining data in category level (*known alignment*). The whole Unknown-Oriented Learning (**UOL**) framework is shown in Fig. 2. For the *true unknown excavation* stage, data in both domains are fed into feature extractor E_1 . Then, a multi-unknown detector W equipped with weight discrepancy constraint L_{weight} is advanced to represent diverse distributions of unseen data and reject these data as unknown category. After that, data gradients are back-propagated and two gradient graphs G_{S2T} , G_{T2T} are built in the *false unknown suppression* stage. Pseudo labels \hat{Y}_T could be obtained through propagating the confidence and rectifying false unknown data. At last, we can align distributions of shared known categories via category-level discriminator $\{D_i\}_{i=1}^K$ and train an extended classifier C to recognize target data.

3.1 True Unknown Excavation

In OSDA, a set of unseen samples are holistically represented by a single unknown category. These unseen samples with various visual characteristics are generally scattered into several clusters in the embedding space, and the corresponding distribution is complicated [23, 1]. Hence, we propose multi-unknown detector (MUD) to explore diverse semantic information of unknown category and excavate more unseen samples. Specifically, the proposed MUD composes of $K + N$ weights \mathbf{w} , i.e. $W = [\mathbf{w}_1, \dots, \mathbf{w}_K, \mathbf{w}_{K+1}, \dots, \mathbf{w}_{K+N}]$, where $\mathbf{w} \in \mathcal{R}^{d \times 1}$, K denotes the number of known categories and N indicates the number of additional weights. After obtaining the feature embedding from feature extractor, $\mathbf{z} = E_1(\mathbf{x}) \in \mathcal{R}^{d \times 1}$, MUD outputs a $K + N$ dimensional logits vector for each sample with the k^{th} output logit $\mathbf{w}_k^\top \mathbf{z}$. Then, the maximum of last N logits is selected to represent the unknown and the softmax is applied to the $K + 1$ dimensional logits vector to produce the posterior probability, $\mathbf{p} = \text{softmax}[\mathbf{w}_1^\top \mathbf{z}, \dots, \mathbf{w}_K^\top \mathbf{z}, \max_{k=K+1, \dots, K+N} \mathbf{w}_k^\top \mathbf{z}]$. In order to identify intrinsic characteristics on unseen categories, weights in W are supposed to have diverse parameters. Therefore, the weight discrepancy constraint is advanced to enforce the divergence of weights via minimizing their cosine similarity,

$$L_{weight}(W) = \sum_{i=1} \sum_{j=1, j \neq i} \frac{\mathbf{w}_i \cdot \mathbf{w}_j}{\|\mathbf{w}_i\| \|\mathbf{w}_j\|}. \quad (1)$$

With this constraint, the introduced weights N for unknown categories will be mutually discrepant, thereby representing various semantics and learning diverse knowledge for unseen categories. To recognize unknown samples in the target

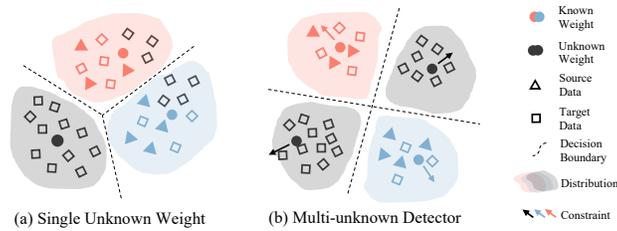


Fig. 3: Comparison between (a) single unknown weight and (b) multi-unknown detector.

domain, an adversarial loss [37] is employed,

$$L_{rej}(x^t) = -\epsilon \log(\mathbf{p}_{K+1}(x^t)) - (1 - \epsilon) \log(1 - \mathbf{p}_{K+1}(x^t)), \quad (2)$$

where ϵ is a pre-defined parameter to balance the adversarial learning, and $\mathbf{p}_{K+1}(\cdot)$ indicates the unknown probability. With gradient reverse [9], this adversarial loss will encourage the unknown probability far away from ϵ , i.e., enhancing it to 1 or suppressing it to 0, and recognize the unknown samples. Moreover, the source knowledge is also utilized to help recognize unknown samples through cross-entropy loss,

$$L_{CE}(x^s, y^s) = -\log(\mathbf{p}_{y^s}(x^s)). \quad (3)$$

To update parameters of feature extractor E_1 and MUD W simultaneously, the gradient reverse layer [9] is introduced to flip the sign of L_{rej} . The object can be formulated as

$$\begin{aligned} E_1^* &= \arg \min_{E_1} L_{CE}(x^s, y^s) - L_{rej}(x^t), \\ W^* &= \arg \min_W L_{CE}(x^s, y^s) + L_{rej}(x^t) + \mu L_{weight}(W). \end{aligned} \quad (4)$$

As shown in Fig. 3, comparing with traditional methods employing a single additional logit of linear classifier to represent the unknown category [37, 23, 1, 8], the proposed MUD and weight discrepancy constraint excavate more unknown samples with various semantic information and achieve inter-class separability.

3.2 False Unknown Suppression

The overwhelming semantic information in unknown category will cause some known samples with similar information to be classified into unknown category. To rectify these false unknown samples, we propose the gradient-graph induced annotation (GIA) module to take advantage of gradient information and build a dense graph structure in sample level to enhance the relationship among known samples, as shown in Fig. 2 (b). To be specific, in the Source-to-Target gradient Graph (S2TAG), the target samples with high confidence are selected as super confident samples with corresponding pseudo labels. Then, based on the

smoothness assumption [48, 13], the intrinsic relationship among target samples in gradient manifold is exploited to propagate the confidence of super confident samples via Target-to-Target grADient Graph (T2TAG). Finally, we acquire refined pseudo labels \hat{Y}_T for the *known alignment* in Sec. 3.3.

Since gradient can reserve the learned knowledge of unseen categories in *True Unknown Excavation* and incorporate information beneficial to recognition, we utilize the gradient via back-propagation to characterize data, i.e. $\hat{\mathbf{g}}^s = \text{Vec}(\frac{\partial}{\partial W} L_{CE}(\mathbf{x}^s, y^s))$, $\hat{\mathbf{g}}^t = \text{Vec}(\frac{\partial}{\partial W} L_{CE}(\mathbf{x}^t, \arg \max_i \mathbf{p}_i))$, where $\text{Vec}(\cdot)$ represents the vectorization. Notably, we only make use of gradient features from the classifier W . Considering the gradient vector with high-dimensional cost high storage and computational resource, kernel-PCA [32, 29] is utilized to determine the low-dimensional gradient manifold and reduce the resource consumption. After applying k-PCA, we obtain vectors $\mathbf{g}^{s/t} \in \mathcal{R}^M$ representing the coordinate, w.r.t. the M principal components in a reproducing kernel Hilbert space, for each sample in the gradient manifold.

S2TAG. To select the target sample with high confidence, we build a bipartite graph $G_{S2T} = \langle V_{S2T}, \mathcal{E}_{S2T} \rangle$ and model the relationship between source and target domains. V_{S2T} includes all samples in source and target domains. The weighted edge matrix encodes the non-negative pairwise similarity between target sample and k -nearest neighbors in source domain, $[\mathcal{E}_{S2T}]_{ij} = [\mathbf{g}_i^t \top \mathbf{g}_j^s]_+$, when $\mathbf{g}_j^s \in \text{NN}_k(\mathbf{g}_i^t)$, otherwise, $[\mathcal{E}_{S2T}]_{ij} = 0$.

Then, we select super confident samples in the target domain via super confidence criteria. For known samples, they are supposed to be similar with source samples of the same category. Their k -neighbors should belong to the same category and the average similarity is large. As for unseen samples, they do not belong to any categories in the source domain. Their k -neighbors should distribute in different categories and the average similarity is small. Based on these two criteria, the super confidence matrix $\tilde{C} \in \mathcal{R}^{n_t \times (K+1)}$ is generated with elements

$$\tilde{C}_{ij}^T = \begin{cases} \mathbf{m}_i, & \text{if } \mathbf{m}_i > \alpha \wedge |N_i| = 1 \wedge j = N_i[0] \\ 1 - \mathbf{m}_i, & \text{if } \mathbf{m}_i < 1 - \alpha \wedge |N_i| > 1 \wedge j = K + 1 \\ 0, & \text{otherwise,} \end{cases} \quad (5)$$

where \mathbf{m} indicates the average similarity vector for target sample, i.e. $\mathbf{m}_i = \text{mean}([\mathcal{E}_{S2T}]_{ij})$ when $\mathbf{g}_j^s \in \text{NN}_k(\mathbf{g}_i^t)$, α denotes the threshold to identify the super confident samples, N_i represents the label set of k -neighbors of each target sample i , and $|\cdot|$ is the cardinality of a set. To this end, the generated confidence matrix encodes the confidence of selected confident samples with super confidence criteria. Moreover, We also generate the corresponding super confident index $\mathbf{s} \in \mathcal{R}^{n_t}$, where 1 indicates the selected sample and vice versa. Through the proposed S2TAG, confident samples in target domain are selected for subsequent propagation.

T2TAG. With selected confident target domain samples, we exploit the intrinsic relationship among target samples in gradient manifold to construct an undirected graph $G_{T2T} = \langle V_{T2T}, \mathcal{A}_{T2T} \rangle$, thereby propagating the confidence in-

formation within the target domain and obtaining the robust pseudo labels. The set of vertices V_{T2T} is composed of all target samples. The data relation in the gradient manifold is encoded using the adjacent matrix \mathcal{A}_{T2T} , and each element is calculated via $[\mathcal{A}_{T2T}]_{ij} = [\mathbf{g}_i^t \top \mathbf{g}_j^t]_+$, when $\mathbf{g}_j^t \in \mathbf{NN}_k(\mathbf{g}_i^t) \wedge j \neq i$, otherwise $[\mathcal{A}_{T2T}]_{ij} = 0$. To ensure the adjacent matrix \mathcal{A}_{T2T} of an undirected graph is symmetric nonnegative with zero diagonal, we have $\mathcal{A}_{T2T} := \mathcal{A}_{T2T} + \mathcal{A}_{T2T}^\top$. Then, the adjacent matrix is normalized symmetrically as $\mathcal{W} = D^{-1/2} \mathcal{A}_{T2T} D^{-1/2}$, where $D = \text{diag}(\mathcal{A}_{T2T} \mathbf{1}_n)$ and $\mathbf{1}_n$ is the all-ones n -vector.

Based on the normalized adjacency matrix \mathcal{W} , we propagate the confidence in matrix \tilde{C} to obtain the confidence matrix for the whole target dataset \hat{C} . With the prior that adjacent samples are supposed to possess comparable confidence [48, 13], an energy function is advanced to measure the propagation result:

$$\mathcal{Q}(\hat{C}) = \text{Tr}(\hat{C}^T (I - \mathcal{W}) \hat{C}) + \lambda \|\mathcal{S} \hat{C} - \tilde{S} \tilde{C}\|_F^2, \quad (6)$$

where $\text{Tr}(\cdot)$ denotes the trace of a squared matrix, $\|\cdot\|_F$ represents the Frobenius norm of matrix, and $S = \text{diag}(\mathbf{s})$. The first term of this energy function describes the smoothness based on the relationship in the manifold, and the second term maintains the corresponding value in the super confident matrix. Through minimizing the energy function, we could obtain the optimal confidence matrix encoding the refined pseudo label for the whole target dataset.

Since this energy function is convex, the global optimal propagation result can be derived when the derivative of energy function is 0, i.e. $\frac{\partial \mathcal{Q}(\hat{C})}{\partial \hat{C}}|_{\hat{C}=\hat{C}^*} = 0$. Thus, the optimal equation is $(I - \mathcal{W} + \lambda S) \hat{C}^* = \lambda \tilde{S} \tilde{C}$. Considering the computational inefficiency of large matrix inversion and the positive definiteness of matrix $(I - \mathcal{W} + \lambda S)$, the conjugate gradient method [2, 13] is applied to solve this linear equation. After propagating, we obtain the pseudo labels with matrix \hat{C}^* to guide the known alignment training, $\hat{Y}_T = \{\hat{y}_i^t | \hat{y}_i^t = \arg \max_j \hat{C}_{ij}^*\}$.

3.3 Known Alignment

After recognizing unknown samples, the OSDA problem turns into CSDA problem. The source and target domains can be aligned in shared label space \mathcal{C}_s without deconstruction from unseen categories. Inspired by previous works [7, 40], the network in this stage is composed of an encoder E_2 , category-level discriminator and extended classifier C . The category-level discriminator have K sub-discriminators $\{D_i\}_{i=1}^K$, each of which aligns the distribution for specific category via a category-level adversarial loss:

$$\begin{aligned} L_{adv}(x^s, x^t) = & \sum_{i=1}^K \mathbb{I}_{i=y^s} \log[D_i(E_2(x^s))] \\ & + \sum_{i=1}^K \mathbb{I}_{i=\hat{y}^t} \log[1 - D_i(E_2(x^t))]. \end{aligned} \quad (7)$$

Moreover, both source and target data are utilized to train the extended classifier via cross entropy, $L_{CE}(x^s, y^s)$ and $L_{CE}(x^t, \hat{y}^t)$ as Eq. 3. During the inference,

only encoder E_2 and extended classifier C need to be reserved. The prediction can be obtained through $C(E_2(x^t))$.

4 Experiments

4.1 Experimental Details

Datasets and Baselines. *Digital Recognition* is a popular benchmark for OSDA with three datasets, i.e. MNIST [19], USPS [12], SVHN [31]. Each dataset contains 10 digitals from 0 to 9. Following the previous work [37], three adaptation tasks is constructed, i.e. SVHN \rightarrow MNIST, MNIST \rightarrow USPS and USPS \rightarrow MNIST.

Office-Home is a challenge domain adaptation benchmark consisting of 15500 images from 65 categories of everyday objects in four domains, i.e. Art (**Ar**), Clipart (**Cp**), Product (**Pr**), and Real-World (**Rw**). We follow the same split set in previous work [23].

Endo-c2k is a new medical diagnosis benchmark related to endoscopy. The source domain is CAD-CAP WCE dataset [6] including 1800 images. The target domain is KID WCE dataset[17] including 2371 images. Three common categories are shared in two domains (normal, inflammatory and vascular), and two categories are exclusive in target domain (polyp and ampulla vater).

We compared the proposed UOL against other state-of-the-art OSDA methods including ATI- λ [17], OSBP [37], AoD [8], STA [23], JPOT [43], ROS[1] and ϵ -OSD [47].

Implementation. For *Digital Recognition*, we use LeNet as backbone. For the other two benchmark, we use the pre-trained ResNet-50 on ImageNet as the backbone. Multi-unknown detector W , extended classifier C and category-level discriminator $\{D_i\}_{i=1}^K$ are fully connected neural networks, and the output dimension of W , C and $\{D_i\}_{i=1}^K$ are $K + N$, $K + 1$ and 1. They are randomly initialized. The additional weights number N is set as 5 for *digital recognition*, 10 for *Office-Home*, 2 for *Endo-c2k*. The hyper-parameters are tuned in Office-Home $Ar \rightarrow Cp$ task and fixed for all other OfficeHome tasks, $\mu = 10^{-3}$, $\alpha = 0.9$, $\lambda = 0.95$. As for the graph part, k -nearest neighbors search is accelerated by Faiss [15], and k is set as 20. Radial basis function is employed for kernel-PCA. We employ SGD optimizer with 0.9 momentum and 10^{-4} decay weight to optimize the network. The learning rate is set as 10^{-3} initially and decreases according to the cosine annealing policy. All experiments are done in GTX 2080Ti GPU. We run each setting 3 times and report the average results.

Evaluation Metrics. Four widely used metrics [37, 1] are employed to measure the performance of all methods, i.e. **OS**: normalized accuracy for all classes, **OS***: normalized accuracy for the known classes only, **Unk**: the accuracy of unknown samples and **HOS**: the harmonic mean of OS^* and UNK . Among all metrics, *HOS* evaluates the performance of OSDA methods more comprehensively. The *HOS* will be high only if the algorithm achieves high performance on both known and unknown category.

Table 1: Classification accuracy (%) on *Digital Recognition* for OSDA. The best and second best results are highlighted by bold and underline separately.

| Method | MNIST→USPS | | | | USPS→MNIST | | | | SVHN→MNIST | | | | Average | | | |
|------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | OS | OS* | Unk | HOS |
| ATI-λ [34] | 86.8 | 89.6 | 73.0 | 80.4 | 82.4 | 81.5 | 86.7 | 84.0 | 67.6 | 66.5 | 73.0 | 69.6 | 78.9 | 79.2 | 77.6 | 78.0 |
| OSBP [37] | 92.1 | <u>94.9</u> | 78.0 | 85.6 | 92.3 | 91.2 | <u>97.6</u> | <u>94.3</u> | 63.0 | 59.1 | 82.3 | 68.8 | 82.4 | 81.7 | 85.9 | 82.9 |
| STA [23] | <u>93.0</u> | <u>94.9</u> | 83.5 | 88.8 | 92.2 | 91.3 | 96.5 | 93.8 | 76.9 | <u>75.4</u> | 84.4 | 79.6 | 87.3 | <u>87.2</u> | 88.1 | <u>87.4</u> |
| AoD [8] | 91.3 | 92.0 | 87.8 | 89.9 | <u>93.1</u> | <u>95.2</u> | 91.7 | 93.4 | 68.6 | 65.5 | 84.3 | 73.7 | 84.3 | 84.2 | 87.9 | 85.7 |
| JPOT [43] | 92.9 | 92.1 | <u>96.9</u> | <u>93.3</u> | 92.4 | 91.2 | 98.4 | 94.2 | <u>79.2</u> | 75.3 | <u>86.7</u> | <u>79.9</u> | <u>88.2</u> | 85.4 | 94.0 | 88.9 |
| UOL | 95.6 | 95.1 | 97.8 | 96.5 | 96.9 | 96.9 | 96.7 | 96.8 | 82.4 | 84.1 | 88.2 | 86.1 | 92.1 | 91.6 | <u>93.5</u> | 93.1 |

Table 2: Recognition accuracy (%) on 12 pairs of source and target domains from *Office-Hone* benchmark including four domains, i.e. Art (*Ar*), Clipart (*Cp*), Product (*Pr*), and Real-World (*Rw*). The best and second best results are highlighted by bold and underline separately.

| Setting | OSBP [37] | | | | AoD [8] | | | | STA [23] | | | | ROS [1] | | | | ε-OSD [47] | | | | UOL | | | |
|-----------------------|-----------|------|------|------|-------------|-------------|------|------|----------|------|-------------|-------------|---------|------|-------------|-------------|-------------|-------------|------|------|-------------|-------------|-------------|-------------|
| | OS | OS* | Unk | HOS | OS | OS* | Unk | HOS | OS | OS* | Unk | HOS | OS | OS* | Unk | HOS | OS | OS* | Unk | HOS | OS | OS* | Unk | HOS |
| <i>Ar</i> → <i>Cp</i> | 50.6 | 50.2 | 61.1 | 55.1 | <u>58.9</u> | <u>59.9</u> | 33.9 | 43.3 | 46.6 | 45.9 | 64.1 | 53.5 | 51.5 | 50.6 | 74.1 | 60.1 | 61.6 | 62.8 | 31.6 | 42.0 | <u>58.9</u> | 58.3 | <u>73.8</u> | 65.2 |
| <i>Ar</i> → <i>Pr</i> | 71.3 | 71.8 | 59.8 | 65.3 | 73.4 | 74.4 | 48.4 | 58.6 | 67.0 | 67.2 | 62.0 | 64.5 | 68.5 | 68.4 | <u>70.3</u> | <u>69.3</u> | <u>76.6</u> | <u>78.3</u> | 34.1 | 47.5 | 78.3 | 78.5 | 72.8 | 75.6 |
| <i>Ar</i> → <i>Rw</i> | 78.8 | 79.3 | 67.5 | 72.9 | 79.2 | 80.2 | 54.2 | 64.7 | 76.2 | 76.7 | 63.7 | 69.6 | 75.9 | 75.8 | <u>77.2</u> | <u>76.5</u> | <u>83.2</u> | <u>85.0</u> | 38.2 | 52.7 | 89.2 | 89.4 | 83.7 | 86.4 |
| <i>Cp</i> → <i>Ar</i> | 59.8 | 59.4 | 70.3 | 64.4 | <u>60.6</u> | <u>61.5</u> | 38.1 | 47.1 | 50.2 | 49.3 | <u>72.7</u> | 58.8 | 54.1 | 53.6 | 65.5 | <u>58.9</u> | 62.2 | 62.8 | 47.2 | 53.9 | 59.6 | 58.8 | 78.9 | 67.4 |
| <i>Cp</i> → <i>Pr</i> | 66.8 | 67.0 | 62.7 | 64.8 | 67.5 | 68.4 | 45.0 | 54.3 | 57.7 | 57.6 | 60.2 | 58.9 | 60.3 | 59.8 | 71.6 | <u>65.2</u> | <u>71.0</u> | 72.2 | 41.0 | 52.3 | 71.7 | <u>71.8</u> | 68.2 | 69.9 |
| <i>Cp</i> → <i>Rw</i> | 71.9 | 72.0 | 69.2 | 70.6 | 74.8 | 75.8 | 49.8 | 60.1 | 64.9 | 65.2 | 57.4 | 61.1 | 65.6 | 65.3 | <u>72.2</u> | <u>68.6</u> | 77.7 | 79.0 | 45.2 | 57.5 | <u>76.0</u> | <u>76.0</u> | 74.1 | 75.1 |
| <i>Pr</i> → <i>Ar</i> | 59.4 | 59.1 | 68.1 | 63.3 | <u>63.8</u> | <u>64.7</u> | 41.3 | 50.4 | 49.5 | 48.4 | 77.0 | 59.4 | 57.6 | 57.3 | <u>64.3</u> | <u>60.6</u> | 64.6 | 65.4 | 44.6 | 53.0 | 64.6 | 64.2 | 74.3 | 68.9 |
| <i>Pr</i> → <i>Cp</i> | 45.3 | 44.5 | 66.3 | 53.3 | <u>58.1</u> | <u>59.0</u> | 35.6 | 44.4 | 42.9 | 40.8 | 95.4 | <u>57.2</u> | 47.5 | 46.5 | 71.2 | 56.3 | 60.0 | 60.8 | 40.0 | 48.3 | 56.6 | 55.6 | <u>81.2</u> | 66.0 |
| <i>Pr</i> → <i>Rw</i> | 76.0 | 76.2 | 71.7 | 73.9 | 77.7 | 78.7 | 52.7 | 63.1 | 76.6 | 77.3 | 59.1 | 67.0 | 71.1 | 70.8 | <u>78.4</u> | <u>74.4</u> | <u>81.5</u> | <u>82.9</u> | 46.5 | 59.6 | 85.1 | 85.2 | 83.2 | 84.2 |
| <i>Rw</i> → <i>Ar</i> | 66.1 | 66.1 | 67.3 | 66.7 | 67.3 | 68.2 | 44.8 | 54.1 | 68.7 | 68.6 | <u>71.2</u> | <u>69.9</u> | 67.1 | 67.0 | 70.8 | 68.8 | <u>70.6</u> | <u>71.6</u> | 45.6 | 55.7 | 75.2 | 75.0 | 79.2 | 77.1 |
| <i>Rw</i> → <i>Cp</i> | 48.6 | 48.0 | 63.0 | 54.5 | 55.8 | 56.7 | 33.3 | 42.0 | 46.0 | 45.4 | <u>61.0</u> | 52.1 | 52.3 | 51.5 | 73.0 | <u>60.4</u> | <u>58.8</u> | <u>59.6</u> | 38.8 | 47.0 | 63.4 | 63.0 | 73.0 | 67.7 |
| <i>Rw</i> → <i>Pr</i> | 76.0 | 76.3 | 68.6 | 72.2 | 77.7 | 78.6 | 55.2 | 64.9 | 73.9 | 74.5 | 58.9 | 65.8 | 72.3 | 72.0 | <u>80.0</u> | <u>75.7</u> | <u>81.3</u> | <u>82.8</u> | 43.8 | 57.3 | 85.5 | 85.3 | 89.3 | 87.2 |
| Average | 64.2 | 64.2 | 66.3 | 65.2 | 67.9 | 68.8 | 45.4 | 54.7 | 60.0 | 59.8 | 65.0 | 62.3 | 62.0 | 61.6 | <u>72.4</u> | <u>66.2</u> | <u>70.8</u> | 71.9 | 43.3 | 54.0 | 72.0 | <u>71.8</u> | 77.6 | 74.2 |

4.2 Results for Benchmarks

Digital Recognition. Experiments are conducted in the digital recognition benchmark. As shown in Table 1, UOL overpasses state-of-the-art methods [34], [37], [23], [8], [43] with 15.1%, 10.2%, 5.7%, 7.4% and 4.2% increments in average *HOS*. These encouraging results demonstrate that the proposed UOL can recognize unknown samples better while achieving high performance in classifying known samples.

Office-Home. We further illustrate the experimental results on all 12 tasks of this benchmark in Table 2. Due to the high openness [23, 1] and large domain gap, we can observe that previous methods achieve poor performance either in unknown recognition or in known classification. For example, ε-OSD [47] only achieves 43.3% unknown category accuracy despite high performance for known classes, leading to an unsatisfactory *HOS* score. On the contrary, UOL achieves competitive result for known categories with 71.8% average *OS** and also superior accuracy 77.6% for unknown samples, which demonstrates the better unknown recognition ability of UOL than ε-OSD [47]. Moreover, UOL surpasses all other OSDA methods with *HOS* metric by a large gap in 12 tasks and possesses superior capability with increments of 26.6%, 19.5%, 11.9%, 8.0% and 20.2%

Table 3: Classification accuracy (%) on *Endo-c2k*.

| Method | Norm. Vasc. | Infl. | Unk | OS | OS* | HOS |
|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Source Only | 31.9 | 53.1 | 31.3 | 0.0 | 29.1 | 38.8 |
| OSBP [37] | 22.8 | 56.8 | <u>63.0</u> | 14.3 | 39.2 | <u>47.5</u> |
| STA [23] | 28.8 | 42.6 | 34.8 | <u>25.4</u> | 32.9 | 35.4 |
| AoD [8] | 20.8 | 55.8 | 65.6 | 22.2 | 41.1 | 47.4 |
| ROS [1] | <u>32.4</u> | <u>59.1</u> | 49.3 | 23.8 | <u>41.2</u> | 46.9 |
| UOL | 38.3 | 59.4 | 56.0 | 30.2 | 46.0 | 51.2 |

Table 4: Ablation performance on the *Ar→Cp* task.

| MUD | WDC | GIA | | | OS | OS* | Unk | HOS |
|-----|-----|-------|-------|-------|------|------|------|------|
| | | Graph | Grad. | k-PCA | | | | |
| | | | | | 53.1 | 53.3 | 48.1 | 50.6 |
| ✓ | | | | | 52.9 | 52.8 | 54.3 | 53.5 |
| ✓ | ✓ | | | | 54.8 | 54.2 | 68.8 | 60.6 |
| ✓ | ✓ | ✓ | | | 56.8 | 56.3 | 69.1 | 62.0 |
| ✓ | ✓ | | ✓ | | 57.7 | 57.1 | 73.1 | 64.1 |
| ✓ | ✓ | ✓ | ✓ | ✓ | 58.9 | 58.3 | 73.8 | 65.2 |

average *HOS* comparing with state-of-the-art methods [37], [8], [23], [1], [47], which reveal that the proposed UOL is robust to the severe domain gap and high openness.

Endo-c2k. Table 3 shows the OSDA results in Endo-c2k benchmark. We firstly evaluate the performance in ‘source only’ setting where only source samples are utilized to train the network, which can be regarded as a lower bound. From the results, we observe that the proposed UOL outperforms all other methods [37], [23], [8], [1] in term of unknown samples recognition with a large gap, i.e. 15.9%, 4.8%, 8.0% and 6.4%. This significant improvement comes from the additional weights representing the unknown category in MUD. Furthermore, the normal category accuracy for OSBP [37], STA [23] and AoD [8] is lower than that for ‘source only’ while that for UOL is not. This shows that UOL can avoid the deconstruction of shared known distributions during alignment.

4.3 Ablation Study

Effectiveness for Multi-Unknown Detector and Weight Discrepancy Constraint. To evaluate the contribution of proposed multi-unknown detector and weight discrepancy constraint, qualitative and quantitative ablation results are shown in Fig. 4 and Table 4, respectively. The additional weight number N is set as 1 in the first row and set as 5 for the second row in Fig. 4. We observe that the distribution of unknown samples often overlap with distributions of known categories for three scenarios in the first row, due to the discrepant visual information inner unknown category. In the second row, the unknown samples are grouped into several clusters with the guidance of multi-unknown weights, and intra-class compactness and inter-class separability is achieved with the help of weight discrepancy constraint. Moreover, comparing the 3rd row and 2nd row with the 1st row in Table 4, the unknown recognition performance increase with 6.2% and 20.7% *Unk*, which verifies that multi-unknown detector (MUD) equipped with weight discrepancy constraint (WDC) can help recognize the unknown samples.

Effectiveness for GIA. We ablate two graph structures, gradient and k-PCA successively to validate the effectiveness of each component in Gradient-graph Induced Annotation. Comparing the three confusion matrices in Fig. 5, lots of off-diagonal elements are non-zero in the 1st matrix, while most of these are

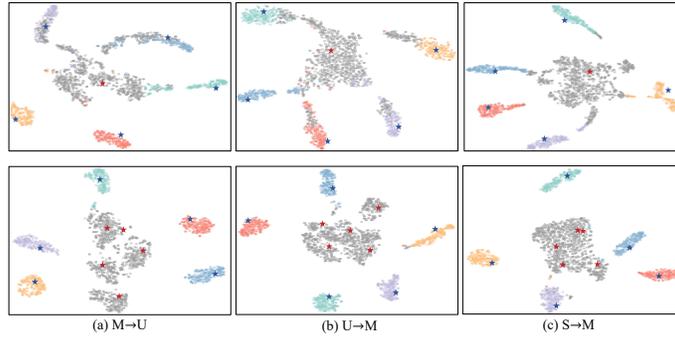


Fig. 4: T-SNE visualization of the target features for the digital recognition benchmark including MNIST (M), USPS (U) and SVHN (S). Gray points indicate the unknown features and points with other colors indicate known features. Red stars represent the classifier weights of unknown category and Blue stars represent the weights of known categories.

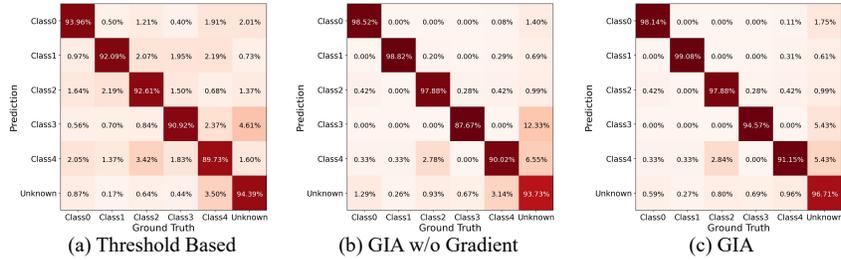


Fig. 5: Confusion matrices of pseudo labels in MNIST→USPS task for (a) threshold base method, (b) GIA without gradient and (c) GIA.

zero in the 2^{nd} and 3^{rd} matrices. This indicates two proposed graph structures could fully utilize the intrinsic relationship of data and suppress noisy pseudo labels. Moreover, the false unknown samples in the 3^{rd} confusion matrix are less than those in 2^{nd} one, which certifies that gradient manifold incorporates the information beneficial to recognition. As for the numerical results in Table 4, it is reported that the unknown accuracy is promoted by 4.0% comparing the 5^{th} row with 4^{th} row, which indicates that gradient can reserve the learned knowledge of unseen categories in the *true unknown excavation*. Comparing the 6^{th} row with the 5^{th} row, the HOS score increases by 1.1% , which validates that k-PCA can learn a appropriate low-dimensional gradient to represent the intrinsic relationship.

Robustness to openness and sensitivity to additional weights number N . In Fig. 6 (a), the relationship between openness and additional weights number N is further discussed. The openness is defined as the $\mathbb{O} = 1 - \frac{|C_s|}{|C_t|}$,

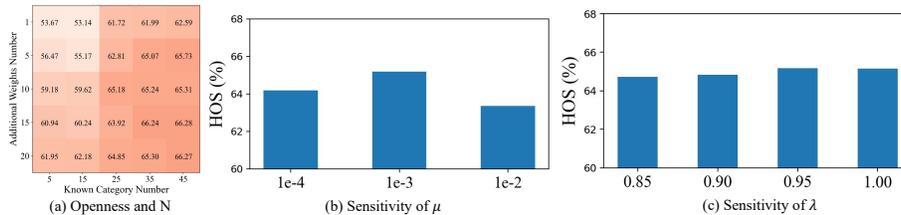


Fig. 6: (a) Ablation study for openness and additional weights number N , (b) Sensitivity analysis of μ and (c) Sensitivity analysis of λ . All experiments are conducted on the $Ar \rightarrow Cp$ task.

where $|\mathcal{C}_s|$ and $|\mathcal{C}_t|$ is the category number is source and target domain. We investigate this setting using $Ar \rightarrow Cp$ task with increasing openness, i.e. 45 known classes $\mathbb{O} = 0.38$, 35 known classes $\mathbb{O} = 0.46$, 25 known classes $\mathbb{O} = 0.62$, 15 known classes $\mathbb{O} = 0.77$ and 5 known classes $\mathbb{O} = 0.92$. For each setting, we conduct experiments with 1, 5, 10, 15 and 20 additional weights number separately. Experimental results represent that UOL framework achieves satisfactory performance in each openness setting with 61.95%, 62.18%, 65.18%, 66.24% and 66.28% HOS , which demonstrates that the proposed method maintains a consistent performance for various openness. Moreover, we observe that HOS scores are stable when the additional weights number is within the interval of [10, 20] in each openness setting. This indicates the results are not sensitive to the additional weight number N within a certain range of [10, 20].

Sensitivity to Coefficients μ and λ . We show the sensitivity analysis for weight discrepancy constraint coefficient μ and super confident matrix coefficient λ in Figure 6 (b) and (c). We vary the value of μ from $1e - 4$ to $1e - 2$ and the value of λ from 0.85 to 1.00. The results show that the scores are stable to these two coefficients within a range. Limited by space, more experiments and implementation details can be found in supplementary material.

5 Conclusion

This paper presents a novel framework **UOL** for open set domain adaptation with a focus on unknown discovery. We propose the multi-unknown detector equipped with weight discrepancy constraint to excavate the diverse unknown knowledge. Meanwhile, gradient-graph induced annotation module is advanced to exploit the intrinsic relationship in gradient manifold and rectify the false unknown samples. After unknown discovery, the remaining samples are aligned in category level. Extensive experiments show that the proposed framework **UOL** performs consistently well on three benchmarks with diverse openness and domain discrepancy. In future, we plan to introduce other prior knowledge such as visual-linguistic embedding to the unknown category.

References

1. Bucci, S., Loghmani, M.R., Tommasi, T.: On the effectiveness of image rotation for open set domain adaptation. In: ECCV. pp. 422–438. Springer (2020)
2. Chandra, S., Kokkinos, I.: Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In: ECCV. pp. 402–418. Springer (2016)
3. Chen, Z., Guo, X., Woo, P.Y., Yuan, Y.: Super-resolution enhanced medical image diagnosis with sample affinity interaction. *IEEE Trans. Med. Imaging* **40**(5), 1377–1389 (2021)
4. Choi, S., Kim, J.T., Choo, J.: Cars can’t fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In: CVPR. pp. 9373–9383 (2020)
5. Cui, S., Wang, S., Zhuo, J., Su, C., Huang, Q., Tian, Q.: Gradually vanishing bridge for adversarial domain adaptation. In: CVPR. pp. 12455–12464 (2020)
6. Dray, X., Li, C., Saurin, J.C., Cholet, F., Rahmi, G., Le Mouel, J., Leandri, C., Lecleire, S., Amiot, X., Delvaux, J.M., et al.: Cad-cap: une base de données française à vocation internationale, pour le développement et la validation d’outils de diagnostic assisté par ordinateur en vidéocapsule endoscopique du grêle. *Endoscopy* **50**(03), 000441 (2018)
7. Du, L., Tan, J., Yang, H., Feng, J., Xue, X., Zheng, Q., Ye, X., Zhang, X.: Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In: CVPR. pp. 982–991 (2019)
8. Feng, Q., Kang, G., Fan, H., Yang, Y.: Attract or distract: Exploit the margin of open set. In: ICCV. pp. 7990–7999 (2019)
9. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML. pp. 1180–1189. PMLR (2015)
10. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. *J Mach Learn Res* **17**(1), 2096–2030 (2016)
11. Guo, X., Yang, C., Liu, Y., Yuan, Y.: Learn to threshold: Thresholdnet with confidence-guided manifold mixup for polyp segmentation. *IEEE Trans. Med. Imaging* **40**(4), 1134–1146 (2020)
12. Hull, J.J.: A database for handwritten text recognition research. *IEEE TPAMI* **16**(5), 550–554 (1994)
13. Iscen, A., Tolias, G., Avrithis, Y., Furon, T., Chum, O.: Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations. In: CVPR. pp. 2077–2086 (2017)
14. Jing, M., Li, J., Zhu, L., Ding, Z., Lu, K., Yang, Y.: Balanced open set domain adaptation via centroid alignment. In: AAAI. vol. 35, pp. 8013–8020 (2021)
15. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. *IEEE Trans. Big Data* (2019)
16. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: CVPR. pp. 4893–4902 (2019)
17. Koulaouzidis, A., Iakovidis, D.K., Yung, D.E., Rondonotti, E., Kopylov, U., Plevris, J.N., Toth, E., Eliakim, A., Johansson, G.W., Marlicz, W., et al.: Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes. *Endosc. Int. Open* **5**(06), E477–E483 (2017)
18. Kwon, G., Prabhushankar, M., Temel, D., AlRegib, G.: Backpropagated gradient representations for anomaly detection. In: ECCV. pp. 206–226. Springer (2020)

19. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**(11), 2278–2324 (1998)
20. Lee, C.Y., Batra, T., Baig, M.H., Ulbricht, D.: Sliced wasserstein discrepancy for unsupervised domain adaptation. In: *CVPR*. pp. 10285–10295 (2019)
21. Li, J., Chen, E., Ding, Z., Zhu, L., Lu, K., Huang, Z.: Cycle-consistent conditional adversarial transfer networks. In: *ACM MM*. pp. 747–755 (2019)
22. Li, J., Chen, E., Ding, Z., Zhu, L., Lu, K., Shen, H.T.: Maximum density divergence for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence* (2020)
23. Liu, H., Cao, Z., Long, M., Wang, J., Yang, Q.: Separate to adapt: Open set domain adaptation via progressive separation. In: *CVPR*. pp. 2927–2936 (2019)
24. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: *CVPR*. pp. 212–220 (2017)
25. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: *ICML*. pp. 97–105. PMLR (2015)
26. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Unsupervised domain adaptation with residual transfer networks. In: *NIPS* (2016)
27. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: *ICML*. pp. 2208–2217. PMLR (2017)
28. Luo, Y., Wang, Z., Huang, Z., Baktashmotlagh, M.: Progressive graph learning for open-set domain adaptation. In: *ICML*. pp. 6468–6478. PMLR (2020)
29. Merrill, N., Olson, C.C.: Unsupervised ensemble-kernel principal component analysis for hyperspectral anomaly detection. In: *CVPR*. pp. 112–113 (2020)
30. Mu, F., Liang, Y., Li, Y.: Gradients as features for deep representation learning. In: *ICLR* (2020), <https://openreview.net/forum?id=BkeoaeHKDS>
31. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)
32. O’Reilly, C., Moessner, K., Nati, M.: Univariate and multivariate time series manifold learning. *Knowl Based Syst* **133**, 1–16 (2017)
33. Pan, Y., Yao, T., Li, Y., Ngo, C.W., Mei, T.: Exploring category-agnostic clusters for open-set domain adaptation. In: *CVPR*. pp. 13867–13875 (2020)
34. Panareda Busto, P., Gall, J.: Open set domain adaptation. In: *ICCV*. pp. 754–763 (2017)
35. Saito, K., Saenko, K.: Ovanet: One-vs-all network for universal domain adaptation. In: *ICCV*. pp. 9000–9009 (2021)
36. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: *CVPR*. pp. 3723–3732 (2018)
37. Saito, K., Yamamoto, S., Ushiku, Y., Harada, T.: Open set domain adaptation by backpropagation. In: *ECCV*. pp. 153–168 (2018)
38. Tang, H., Chen, K., Jia, K.: Unsupervised domain adaptation via structurally regularized deep clustering. In: *CVPR*. pp. 8725–8735 (2020)
39. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: *CVPR*. pp. 7167–7176 (2017)
40. VS, V., Gupta, V., Oza, P., Sindagi, V.A., Patel, V.M.: Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In: *CVPR*. pp. 4516–4526 (2021)
41. Wang, X., Li, L., Ye, W., Long, M., Wang, J.: Transferable attention for domain adaptation. In: *AAAI*. vol. 33, pp. 5345–5352 (2019)
42. Xie, S., Zheng, Z., Chen, L., Chen, C.: Learning semantic representations for unsupervised domain adaptation. In: *ICML*. pp. 5423–5432. PMLR (2018)

43. Xu, R., Liu, P., Zhang, Y., Cai, F., Wang, J., Liang, S., Ying, H., Yin, J.: Joint partial optimal transport for open set domain adaptation. In: IJCAI. pp. 2540–2546 (2020)
44. Zellinger, W., Grubinger, T., Lughofer, E., Natschläger, T., Saminger-Platz, S.: Central moment discrepancy (cmd) for domain-invariant representation learning. arXiv preprint arXiv:1702.08811 (2017)
45. Zhang, Y., Tang, H., Jia, K., Tan, M.: Domain-symmetric networks for adversarial domain adaptation. In: CVPR. pp. 5031–5040 (2019)
46. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR. pp. 6881–6890 (2021)
47. Zhong, L., Fang, Z., Liu, F., Yuan, B., Zhang, G., Lu, J.: Bridging the theoretical bound and deep algorithms for open set domain adaptation. *IEEE Trans Neural Netw Learn* (2021)
48. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: NeurIPS. pp. 321–328 (2004)