Dual Structural Knowledge Interaction for Domain Adaptation

Yukun Zuo, *Graduate Student Member, IEEE*, Hantao Yao, *Member, IEEE*, Liansheng Zhuang, *Member, IEEE*, and Changsheng Xu, *Fellow, IEEE*

***Abstract*—Domain adaptation aims to transfer knowledge from a label-rich source domain to an unlabeled target domain. A common strategy is to assign pseudo-labels to unlabeled target samples for performing representation learning. However, most existing methods only apply the source-guided classifier to generate the source-biased pseudo-labels for self-training, leading to biased target representations. Moreover, the generated pseudo-labels ignore the manifold assumption that neighboring samples are likely to have the same labels. To address the above problem, we formulate a novel structural knowledge to assign target-oriented and manifold-guided pseudo-labels for unlabeled target samples. The structural knowledge consists of cluster-based knowledge and locality-based knowledge. The cluster-based knowledge denotes the label consistency between the target samples and the non-parametric target cluster centers, making the pseudo-labels target-oriented. The locality-based knowledge constrains the target sample and its neighbors to satisfy the manifold assumption. As the neighbors contain the source and target samples, the source and target locality-based knowledge are utilized to boost the descriptions. With the structural knowledge, we propose a novel Dual Structural Knowledge Interaction (DSKI) framework for domain adaptation. For generating aligned and discriminative features, knowledge consistency constraint and instance mutual constraint are proposed in DSKI. Evaluations on three benchmarks demonstrate the effectiveness of the Dual Structural Knowledge Interaction, *e.g.,* 74.9%, 87.7%, and 90.8% for Office-Home, VisDa-2017, and Office-31, respectively.**

***Index Terms*—Domain adaptation, structural knowledge, dual structural knowledge interaction.**

Manuscript received 5 July 2022; revised 14 December 2022; accepted 10 February 2023. Date of publication 21 February 2023; date of current version 15 December 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0112200, in part by the National Natural Science Foundation of China under Grants 62036012, 61721004, U21B2044, U20B2070, and 61976199, and in part by Beijing Natural Science Foundation under Grants L201001 and 4222039. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. De-Nian Yang. *(Corresponding author: Changsheng Xu.)*

Yukun Zuo and Liansheng Zhuang are with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: zykpy@mail.ustc.edu.cn; lszhuang@ustc.edu.cn).

Hantao Yao is with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yao@nlpr.ia.ac.cn).

Changsheng Xu is with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: csxu@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TMM.2023.3245420

I. Introduction

DOMAIN adaptation intends to learn a well-performing model from a fully-labeled source domain and applies it to an unlabeled target domain. Since domain adaptation obtains the satisfactory performance in target domain without the demand of the target labels, it has received much attention in real-world scenarios. The critical issue of domain adaptation is how to align the source and target representations to reduce domain bias. Recently, a lot of methods have been proposed to explicitly align data distribution, *e.g.,* statistical-based methods [1], [2], [3], [4] and adversarial-based methods [5], [6], [7], [8], [9].

Since pseudo-labeling methods [10], [11], [12] are effectively used for semi-supervised learning, a lot of methods [13], [14], [15], [16] apply the pseudo-labeling for domain adaptation, *i.e.,* assigning pseudo-labels to the unlabeled target samples for self-training. These methods obtain pseudo-labels by picking up the maximum class probability based on the *implicit source cluster-based prior*, whose predicted probability is denoted as the similarity between its feature and the implicit source cluster centers, as shown in Fig. 1(a). For example, DSBN [13] and PFAN [15] treat the parameters of the classifier inferred with the labeled source samples as the implicit source cluster centers to predict the target probability for pseudo-labeling. However, the pseudo-labels generated by merely considering the implicit source cluster-based prior contain many incorrect labels for the following two reasons: 1) the implicit source cluster centers are biased against the target samples due to the domain gap between the source and target domains; 2) the implicit source cluster-based prior ignores the manifold assumption that neighboring samples are likely to have the same labels.

To address the above problem, we formulate a novel structural knowledge to obtain the target-oriented and manifold-guided pseudo-labels for unlabeled target samples, consisting of *cluster-based knowledge* and *locality-based knowledge*. The cluster-based knowledge denotes *the label consistency between the target sample and the target cluster center.* To reduce the bias of the cluster center, we formulate non-parametric centers by considering the target features and their predicted probabilities. Moreover, the target sample and its neighbors are likely to have the similar labels based on the manifold assumption. Therefore, the locality-based knowledge is constructed by considering the correlation between the target sample and its selected neighbors, *i.e.,* the higher similarity, the higher confidence with the same labels. Furthermore, as the neighbors contain the source and

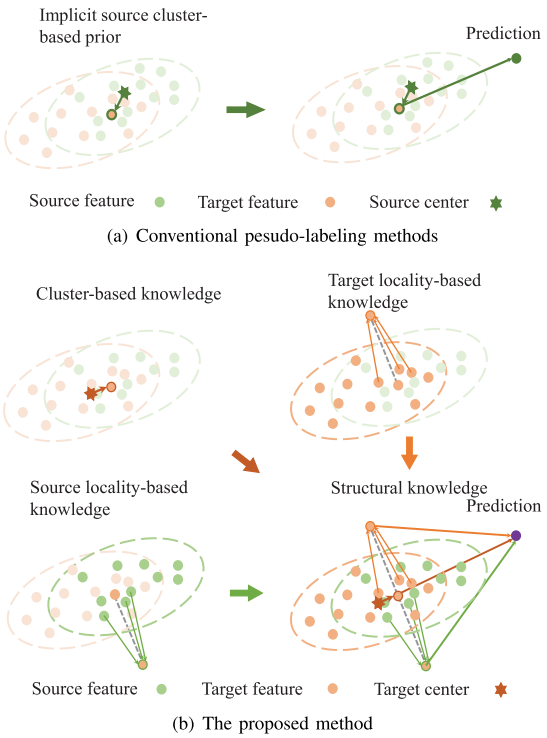


Fig. 1. (a) Conventional pseudo-labeling methods only consider the similarity between the target sample and implicit source cluster center. (b) The proposed method aggregates the cluster-based knowledge, target locality-based knowledge, and source locality-based knowledge together to obtain structural knowledge for pseudo-labeling. The nearest center is presented for convenience.

target samples simultaneously, the source and target locality-based knowledge can provide valuable clues for predicting the target probability. Therefore, the locality-based knowledge consists of the source and target locality-based knowledge. By considering cluster-based knowledge and locality-based knowledge jointly, using the structural knowledge can obtain a robust probability of each target sample, as shown in Fig. 1(b).

With the structural knowledge, we propose a novel Dual Structural Knowledge Interaction (DSKI) framework for domain adaptation, as shown in Fig. 2. To increase the diversity of structural knowledge of unlabeled target samples, we apply the weak augmentation and strong augmentation strategies to augment the target samples for inferring the weak and strong spaces. Furthermore, two types of structural knowledge are aggregated and interacted between the weak and strong feature spaces. Specifically, for constructing the cluster-based knowledge, we utilize Memory Bank [17] to store the non-parametric class centers of the target domain considering the probabilities produced by the classifier. The label consistency is constrained between each target sample and the target class centers. Moreover, for constructing the source and target locality-based knowledge, we also utilize Memory Bank [17] to store the source and target descriptions of the source samples and the augmented target samples, and contrain each target sample and its neighbors to have similar label. With the memory bank of each space, we aggregate the cluster-based knowledge and locality-based knowledge to obtain the weak and strong structural knowledges. Since both structural knowledges describe the same target sample, the knowledge consistency constraint is applied to align two types of structural knowledge for target representation learning, *i.e.,* the weak (strong) structural knowledge is used to optimize the strong (weak) feature space. Besides, we also apply the instance mutual constraint to promote the discrimination of target representations.

The contributions of this work are summarized as follows:

- We introduce a novel structural knowledge to capture the cluster-based knowledge and locality-based knowledge for assigning high-quality pseudo-labels to the target samples.
- By considering the introduced structural knowledge, a novel Dual Structural Knowledge Interaction (DSKI) framework is proposed for domain adaptation.
- Evaluations on three benchmarks demonstrate the effectiveness of the introduced structural knowledge and the Dual Structural Knowledge Interaction, *e.g.,* obtaining the mean accuracy of 74.9%, 87.7%, and 90.8% for Office-Home, VisDa-2017, and Office-31, respectively.

II. Related Work

Recently, a lot of domain adaptation methods [18], [19], [20], [21], [22], [23], [24] have been proposed to reduce the domain gap, which can be divided into three categories: statistical-based methods, adversarial learning methods and pseudo-labeling methods. Note that the proposed method belongs to pseudo-labeling methods.

A. Statistical-Based Methods

Since the data distributions between the source and target domains are different, various distance metrics of domain discrepancy are adopted for feature alignment. For example, some methods utilize Maximum Mean Discrepancy (MMD) [3], [25] as a measure of domain gap to align the means of the source and target features. JAN [1] proposes Joint Maximum Mean Discrepancy (JMMD) to align the joint distributions of multiple domain-specific layers between source and target domains. CORAL [26] aligns the second-order statistics of the source and target distributions, and CMD [27] matches the higher-order central moments of probability distributions by means of order-wise moment differences. CAN [28] proposes a new Contrastive Domain Discrepancy (CDD) to model the intra-class domain discrepancy and the inter-class domain discrepancy considering the class information. Apart from MMD, Wasserstein distance [29], [30], [31], [32], [33] is also widely used for feature alignment. Wasserstein distance is originally from optimal transport [34] problem which seeks an optimal way to transport material from mines to factories, but it is also utilized to measure the discrepancy between source and target domains. OPDA [29] firstly proposes to align the data distribution in the source and target domains with a regularized unsupervised optimal transportation model. JDOT [33] estimates the joint feature/label space distribution of the source and target domains with optimal transport. SWD [32] captures the dissimilarity between the outputs of task-specific classifiers with Sliced Wasserstein Discrepancy. Although these methods are helpful to reduce domain discrepancy, they fail to exploit the label information of outputs for representation learning and obtain inferior results. GPDA [35] executes graph dual regularization in the matrix factorization framework to learn the discriminative and domain-invariant features, while preserving both the statistical properties and geometrical structures of the

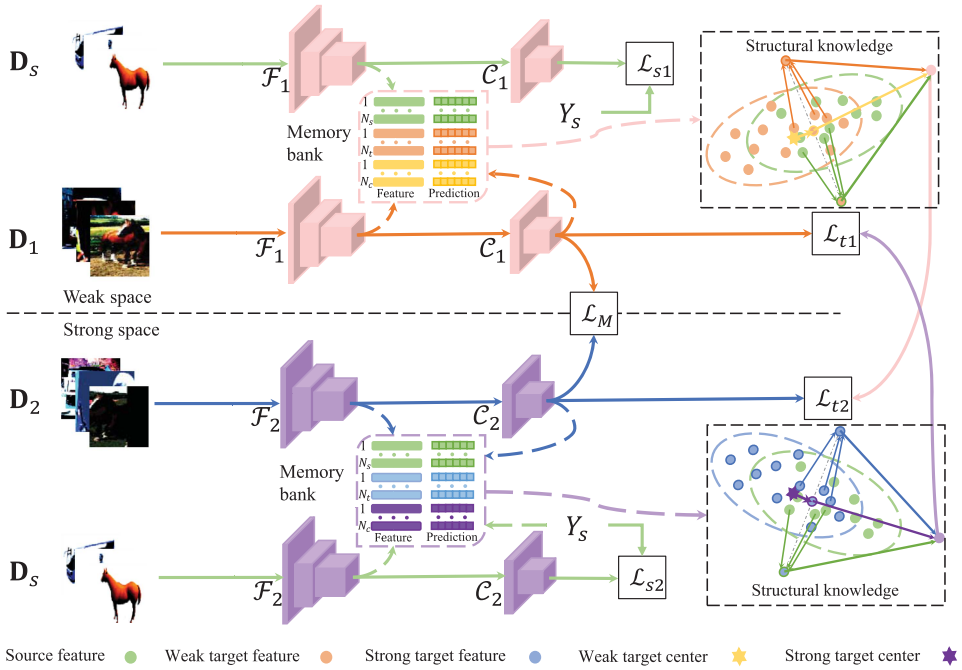


Fig. 2. The framework of Dual Structural Knowledge Interaction(DSKI). DSKI infers the weak-augmented target samples $\mathbf{D}_1$ and strong-augmented target samples $\mathbf{D}_2$ in independent networks to form the weak and strong embedding spaces, respectively. Given the source samples $\mathbf{D}_s$, weak-augmented target samples $\mathbf{D}_1$ and strong-augmented target samples $\mathbf{D}_2$, each space aggregates the knowledge via constructing memory bank to obtain structural knowledge, and propagates its structural label information to its peer space. Furthermore, the mutual information between two embedding space is maximized for discriminative representation learning.

original data. Specifically, graph dual regularization constrains similar examples or features should have similar embeddings, focusing on the consistency of manifold structures. Different from GPDA, the manifold assumption in the proposed DSKI refers that neighboring samples are likely to have the same labels, concentrating on the label consistency in manifold space of the source and target domains. Moreover, GPDA utilizes class centers for domain alignment, while the proposed DSKI considers class centers for obtaining reliable pseudo-labels.

B. Adversarial Learning Methods

As Generative Adversarial Network (GAN) [36] has achieved great success in image generation, many adversarial learning methods [37], [38], [39], [40] have been proposed for feature alignment in domain adaptation recently. DANN [5] adopts the domain classifier to distinguish the source and target features, and utilizes the feature extractor to confuse the domain classifier for learning indistinguishable features. CDAN [41] presents conditional adversarial domain adaptation utilizing the discriminative information of the classifier predictions. MSTN [6] aligns the class centroids in the source and target domains to learn semantic representations for unlabeled target samples. GVB [42] proposes a gradually vanishing bridge mechanism on both generator and discriminator to model domain-invariant and domain-specific parts in the representations. Another type of adversarial methods [43], [44], [45], [46] utilizes multiple different task-specific classifiers as a domain classifier. MCD [43] maximizes the discrepancy between the outputs of two classifiers to detect target samples that are far from the support of the source, and makes the feature extractor generate target features near the support to minimize the discrepancy. STAR [45] models classifier as a Gaussian distribution with its variance representing the inter-classifier discrepancy rather than a weight vector, thus an arbitrary number of classifiers can be used. BCDM [46] designs a novel classifier determinacy disparity formulating classifier discrepancy as the class relevance of distinct target predictions to generate discriminative representations. However, the training of adversarial learning is unstable and adversarial learning methods ignore the connections between samples, which results in degraded performance for domain adaptation.

C. Pseudo-Labeling Methods

Due to the marvelous performance of pseudo-labeling methods [10], [11], [12] in semi-supervised learning, pseudo-labeling methods [13], [15], [47], [48], [49], [50] in domain adaptation have attracted considerable attention recently. JAD [47] gains the pseudo-labels via the classifier trained on the source samples, and infers an improved classification model with the pseudo-labeled target samples together with source samples. MADA [51] obtains the conditional probability of each class for target samples, resulting in soft pseudo-labels. However, the classifier inferred with the labeled source samples cannot produce robust pseudo-labels for target samples because of the domain gap between the source and target domains. NRC [52] exploits local structure of target data, *i.e.*, the local neighbors, reciprocal neighbors, and the expanded neighborhood, to encourage label consistency among data. SND [53] computes the entropy of the similarity distribution between target samples to measure the density of soft neighborhoods, which serves as

an unsupervised validation criterion. However, these two methods do not explicitly utilize the information of neighbors to obtain pseudo-labels for supervised learning. Different from the above methods, the proposed DSKI explicitly adopts the predicted probabilities of target neighbors for obtaining pseudo-labels. Furthermore, NRC and SND ignore the valuable neighborhood structure in the source domain. Nevertheless, the proposed DSKI jointly considers the locality-based knowledge in the source domain and target domain for reliable pseudo-labels. SFDA-DE [54] obtains target centers with spherical k-means, and gains pseudo-labels via the label of the nearest target center. Similar to SFDA-DE, CDS [55] first utilizes k-means to obtain target centers, and acquires pseudo-labels with the nearest target center for cross-domain person re-identification. Different from SFDA-DE and CDS, the proposed DSKI considers the distances of each target sample with all the target centers to obtain cluster-based knowledge. CCAN [56] adopts Graph Convolutional Network to encode data structure information for obtaining GCN feature, then GCN features and CNN features are concatenated for domain alignment as well as class centroid alignment. AdaGraph [57] focuses on predictive domain adaptation scenario, which learns to generalize from annotated source images plus unlabeled samples with associated metadata from auxiliary domains to unlabeled target domain. It builds a graph to describe the dependencies among different domains, and exploits the connection between the target domain and auxiliary domains with constructed graph to regress the target model at test time. Although GCAN and AdaGraph build graph to solve domain adaptation problem, they ignore the locality-based knowledge, *i.e.*, each target sample and its neighbors are likely to have similar labels. Different from GCAN and AdaGraph, the proposed DSKI constructs graphs to exploit the underlying source and target locality-based knowledge for reliable pseudo-labels of unlabeled target samples.

The methods most similar to ours are SRDC [49] and ATDOC [50]. SRDC [49] uncovers the intrinsic target discrimination via discriminative clustering of target data, and implements structure source regularization to implicitly achieve feature alignment. However, SRDC only considers the cluster-based knowledge and ignores the locality-based knowledge in the source and target domains. Moreover, ATDOC develops two types of non-parametric classifiers, *i.e.*, the nearest centroid classifier, and neighborhood aggregation classifier, to improve the quality of pseudo-labels. However, ATDOC treats these two classifiers separately, and ignores the knowledge interaction between two domains. Unlike SRDC and ATDOC, the proposed method jointly considers the cluster-based knowledge contained in target cluster centers and the locality-based knowledge contained in nearby source and target samples. Furthermore, it interacts the knowledge between two different feature spaces for representation learning.

III. Structural Knowledge

A. Problem Formulation

Domain adaptation focuses on transferring knowledge from a labeled source dataset to an unlabeled target one. Formally, the source and target datasets are defined as $\mathbf{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and $\mathbf{D}_t = \{(x_i^t)\}_{i=1}^{N_t}$, where N_s and N_t represent the number of source and target samples, respectively. Furthermore, there are the same N_c categories in the source and target domains. The goal of deep domain adaptation is to infer an unbiased feature extractor $\mathcal{F}$ and classifier $\mathcal{C}$. Given the target sample x, the pseudo-labeling methods firstly apply the feature extractor $\mathcal{F}$ to extract its description $\mathbf{f} = \mathcal{F}(x)$. Then they utilize the classifier $\mathcal{C}$ to predict its probability $\mathbf{p}$ and select the category with the maximum predicted probability as pseudo-label. After that, the supervised classification loss $\mathcal{L}$ is used to optimize the $\mathcal{F}$ and classifier $\mathcal{C}$ on target dataset $\mathbf{D}_t$ with (2),

$$\mathbf{p} = \mathcal{C}(\mathcal{F}(x)), \qquad \hat{y} = \arg\max \mathbf{p}, \tag{1}$$

$$\mathcal{L} = -\frac{1}{N_t} \sum_{x \in \mathbf{D}_t} \hat{y} log \mathbf{p}, \tag{2}$$

where $\mathbf{p}$ and $\hat{y}$ denote the predicted probability and pseudo-label for target sample x.

However, the above-mentioned pseudo-labeling methods have two disadvantages for domain adaptation: 1) the classifier $\mathcal{C}$ inferred with the labeled source samples is biased to the source samples because of the existence of domain gap; 2) only the target sample x itself is considered to generate pseudo-label, leading to that the target sample x and its neighbors might do not satisfy the manifold assumption. To address these problems, we formulate a novel structural knowledge to aggregate the cluster-based knowledge and locality-based knowledge of each target sample for pseudo-labeling. Specifically, the cluster-based knowledge $\mathcal{P}^c(\mathcal{F}(x), \mathbb{C})$ represents the similarity between the target feature $\mathcal{F}(x)$ and the target cluster sets $\mathbb{C}$. The locality-based knowledge $\mathcal{P}^l(\mathcal{F}(x), \mathbb{N}(x))$ constrains the target feature $\mathcal{F}(x)$ and its neighbors $\mathbb{N}(x)$ to satisfy the manifold assumption. By considering these two kinds of knowledge, the pseudo-label of the target sample x can be obtained by (3),

$$\mathcal{P}(x) = \beta \mathcal{P}^c(\mathcal{F}(x), \mathbb{C}) + (1-\beta)\mathcal{P}^l(\mathcal{F}(x), \mathbb{N}(x)), \tag{3}$$

where $\mathcal{P}(x)$ denotes the final class probability of the target sample x, and $\beta \in [0, 1]$ is a trade-off parameter. In the following, we give detailed descriptions about $\mathcal{P}^c(\mathcal{F}(x), \mathbb{C})$ and $\mathcal{P}^l(\mathcal{F}(x), \mathbb{N}(x))$.

B. Cluster-Based Knowledge

The cluster-based knowledge aims to model the correlation between the target feature and the target cluster center, *i.e.*, the higher the similarity, the higher the consistency of the category probability. Formally, the probability of the target sample x is a combination of the center category probabilities, *e.g.*, $\mathcal{P}^c(\mathbf{f}, \mathbb{C}) \in \mathbb{R}^{N_c \times 1}$ denotes the cluster-based probability, where $\mathbf{f} = \mathcal{F}(x)$ is the target feature and $\mathbb{C} = \{(\mathbf{c}_i, \mathbf{p}_i^c)\}_{i=1}^{N_c}$ denotes the set of target centers. Specifically, $\mathbf{c}_i$ and $\mathbf{p}_i^c$ are the feature and probability of the i-th center, respectively. To transfer the probability information from the target center $(\mathbf{c}_i, \mathbf{p}_i^c)$ to the target feature $\mathbf{f}$, we construct a graph $G = (V, E)$ to exploit the underlying cluster-based knowledge. The node set $V = \{v_i\}_{i=0}^{N_c}$ consists of the given target feature $\mathbf{f}$ and N_c centers, where $v_i = (\mathbf{c}_i, \mathbf{p}_i^c)$ contains the node feature $\mathbf{c}_i$ and its probability $\mathbf{p}_i^c$. The set of edges $E = \{e_i\}_{i=1}^{N_c}$ represents the connections

between the target feature $\mathbf{f}$ and all N_c centers, where e_i is the similarity function to measure the target feature $\mathbf{f}$ and the center feature $\mathbf{c}_i$ computed with (4),

$$e_i = \frac{\exp(\mathbf{f}^\top \mathbf{c}_i)}{\sum_{j=1}^{N_c} \exp(\mathbf{f}^\top \mathbf{c}_j)}. \tag{4}$$

Based on the graph G, the probability of the target feature is the aggregation of all nodes with (5),

$$\mathcal{P}^c(\mathcal{F}(x), \mathbb{C}) = \sum_{i=1}^{N_c} e_i \mathbf{p}_i^c = \sum_{i=1}^{N_c} \frac{\exp(\mathbf{f}^\top \mathbf{c}_i)}{\sum_{j=1}^{N_c} \exp(\mathbf{f}^\top \mathbf{c}_j)} \mathbf{p}_i^c, \tag{5}$$

where $\mathbf{f} = \mathcal{F}(x)$ and $\mathbb{C} = \{(\mathbf{c}_i, \mathbf{p}_i^c)\}_{i=1}^{N_c}$.

To optimize (5), the rest problem is how to obtain the center sets $\mathbb{C} = \{(\mathbf{c}_i, \mathbf{p}_i^c)\}_{i=1}^{N_c}$. The center feature is obtained by fusing the current centers $\mathbf{c}_i$ and the new center $\hat{\mathbf{c}}_i$ with (7),

$$\mathbf{c}_i = (1-\tau)\mathbf{c}_i + \tau \hat{\mathbf{c}}_i, \tag{6}$$

where τ is a smoothing coefficient hyperparameter set to 0.1 by default. The new center $\hat{\mathbf{c}}_i$ is generated by applying the average of the target features predicted to the i-th class.

$$\hat{\mathbf{c}}_i = \frac{\sum_{k=1}^{N_t} \mathbf{p}_{k,i} \mathbf{f}_k}{\sum_{k=1}^{N_t} \mathbf{p}_{k,i}}, \tag{7}$$

where $\mathbf{f}_k$ represents the feature of the k-th target sample, $\mathbf{p}_{k,i}$ denotes the probability of the k-th target samples predicted to the i-th class.

For the center probability $\mathbf{p}_i^c$, there are two situations: soft probability and hard probability. The hard probability is a one-hot vector for describing the center belonging to each class, and the soft probability is obtained by feeding the target center feature $\mathbf{c}_i$ into the classifier $\mathcal{C}$. Here we adopt hard probability as default.

C. Locality-Based Knowledge

Different from cluster-based knowledge, locality-based knowledge aims to model the relationship between the target sample and its neighbors for satisfying the manifold assumption, *i.e.,* the higher similarity between the target sample and its neighbors, the higher confidence with the same labels. For convenience, we define the feature of a given target sample x as $\mathbf{f} = \mathcal{F}(x)$. Therefore, the locality-based knowledge $\mathcal{P}^l(\mathcal{F}(x), \mathbb{N}(x))$ can also be denoted as $\mathcal{P}^l(\mathbf{f}, \mathbb{N}(x))$. As the neighbors contain the source and target samples simultaneously, the locality-based knowledge $\mathcal{P}^l(\mathbf{f}, \mathbb{N}(x))$ is a combination of the target locality-based knowledge $\mathcal{P}^t(\mathbf{f}, \mathbb{N}^t(x))$ and the source locality-based knowledge $\mathcal{P}^s(\mathbf{f}, \mathbb{N}^s(x))$,

$$\mathcal{P}^l(\mathbf{f}, \mathbb{N}(x)) = \gamma \mathcal{P}^t(\mathbf{f}, \mathbb{N}^t(x)) + (1-\gamma)\mathcal{P}^s(\mathbf{f}, \mathbb{N}^s(x)), \tag{8}$$

where $\gamma \in [0, 1]$ is a trade-off parameter.

As the way to obtain the source locality-based knowledge $\mathcal{P}^s(\mathbf{f}, \mathbb{N}^s(x))$ is similar to that of the target locality-based knowledge $\mathcal{P}^t(\mathbf{f}, \mathbb{N}^t(x))$, we thus introduce the $\mathcal{P}^t(\mathbf{f}, \mathbb{N}^t(x))$ for simplicity. For the target locality-based knowledge $\mathcal{P}^t(\mathbf{f}, \mathbb{N}^t(x))$, $\mathbb{N}^t(x) = \{(\mathbf{f}_i^t, \mathbf{p}_i^t)\}_{i=1}^K$ denotes the features and predicted probabilities of the neighbors, where the probabilities are generated by the classifiers. Given the target sample x, its neighbors are generated by selecting the K-nearest neighbors based on feature similarity. Similar to the cluster-based knowledge, we construct a graph $G^t = (V^t, E^t)$ to exploit the underlying target locality-based knowledge, where $V^t = \{v_i^t\}_{i=0}^K$ includes the given target sample and K neighboring target samples, and $E^t = \{e_i^t\}_{i=1}^K$ comprises the connections between the target feature and K neighboring target features. Each node $v_i^t = (\mathbf{f}_i^t, \mathbf{p}_i^t)$ contains the node feature $\mathbf{f}_i^t$ and its probability $\mathbf{p}_i^t$, and the edge e_i^t is the distance between the target feature $\mathbf{f}$ and the neighboring target feature $\mathbf{f}_i^t$,

$$e_i^t = \frac{\exp(\mathbf{f}^\top \mathbf{f}_i^t)}{\sum_{j=1}^{K} \exp(\mathbf{f}^\top \mathbf{f}_j^t)}. \tag{9}$$

According to the graph G^t, the target locality-based probability is aggregated with (10),

$$\mathcal{P}^t(\mathbf{f}, \mathbb{N}^t(x)) = \sum_{i=1}^{K} e_i^t \mathbf{p}_i^t = \sum_{i=1}^{K} \frac{\exp(\mathbf{f}^\top \mathbf{f}_i^t)}{\sum_{j=1}^{K} \exp(\mathbf{f}^\top \mathbf{f}_j^t)} \mathbf{p}_i^t. \tag{10}$$

Similar to the target locality-based knowledge, the source locality-based knowledge $\mathcal{P}^s(\mathbf{f}, \mathbb{N}^s(x))$ is generated with (11),

$$\mathcal{P}^s(\mathbf{f}, \mathbb{N}^s(x)) = \sum_{i=1}^{K} e_i^s \mathbf{p}_i^s = \sum_{i=1}^{K} \frac{\exp(\mathbf{f}^\top \mathbf{f}_i^s)}{\sum_{j=1}^{K} \exp(\mathbf{f}^\top \mathbf{f}_j^s)} \mathbf{p}_i^s, \tag{11}$$

where $\mathbf{f}_i^s$ represents the source node feature, and $\mathbf{p}_i^s$ depicts the source predicted probability. Since the source labels are available, the source predicted probabilities $\{\mathbf{p}_i^s\}_{i=1}^{N_s}$ are obtained by label smoothing strategy [58] with (12),

$$\mathbf{p}_{i,j}^s = \begin{cases} \frac{\epsilon}{N_c} & \text{if } \ j \neq y_i^s, \\ 1 - \epsilon + \frac{\epsilon}{N_c} & \text{if } \ j = y_i^s, \end{cases} \tag{12}$$

where $\mathbf{p}_{i,j}^s$ indicates the j-th class probability of $\mathbf{p}_i^s$, ϵ denotes the hyperparameter for smoothing, y_i^s represents the corresponding ground-truth label, and N_c is the number of the classes.

D. Aggregation

For the unlabeled target samples x, the structural knowledge $\mathcal{P}(x)$ is the combination of $\mathcal{P}^c$, $\mathcal{P}^t$ and $\mathcal{P}^s$ for optimization,

$$\min_{\mathcal{P}(x) \in \mathbb{R}^{N_c}} \mathcal{L}(\mathcal{P}(x)) = \beta \|\mathcal{P}(x) - \mathcal{P}^c(x)\|_2^2 + (1-\beta)\left[\gamma \|\mathcal{P}(x) - \mathcal{P}^t(x)\|_2^2 + (1-\gamma)\|\mathcal{P}(x) - \mathcal{P}^s(x)\|_2^2\right], \tag{13}$$

where $\|\mathcal{P}(x) - \mathcal{P}^c(x)\|_2^2$ constrains $\mathcal{P}(x)$ to be close to the cluster-based knowledge $\mathcal{P}^c(x)$, $\|\mathcal{P}(x) - \mathcal{P}^t(x)\|_2^2$ constrains that $\mathcal{P}(x)$ should not change too much with the target locality-based probability $\mathcal{P}^t(x)$, $\|\mathcal{P}(x) - \mathcal{P}^s(x)\|_2^2$ means that $\mathcal{P}(x)$ has a similar value with the source locality-based probability $\mathcal{P}^s(x)$. Note that $\mathcal{P}^c(x)$, $\mathcal{P}^t(x)$, and $\mathcal{P}^s(x)$ are the abbreviation of $\mathcal{P}^c(\mathbf{f}, \mathbb{C})$, $\mathcal{P}^t(\mathbf{f}, \mathbb{N}^t(x))$, and $\mathcal{P}^s(\mathbf{f}, \mathbb{N}^s(x))$, respectively. $\beta \in [0, 1]$ and $\gamma \in [0, 1]$ are two hyperparameters.

To optimize the (13), we observe the objective function is (convex) quadratic. Since any locally optimal point of a convex optimization problem is also global optimal [59], we conduct $\nabla\mathcal{L}(\mathcal{P}^*(x)) = 0$ to obtain the minimizer $\mathcal{P}^*(x)$ of $\mathcal{L}(\mathcal{P}(x))$:

$$\begin{aligned}
&\nabla\mathcal{L}(\mathcal{P}^*(x)) \\
&= 2\beta(\mathcal{P}^*(x) - \mathcal{P}^c(x)) + 2(1-\beta)[\gamma(\mathcal{P}^*(x) - \mathcal{P}^t(x)) \\
&\quad + (1-\gamma)(\mathcal{P}^*(x) - \mathcal{P}^s(x))] \\
&= 2\mathcal{P}^*(x) - 2\beta\mathcal{P}^c(x) - 2(1-\beta)[\gamma\mathcal{P}^t(x) + (1-\gamma)\mathcal{P}^s(x)] \\
&= 0
\end{aligned} \tag{14}$$

Therefore, the minimizer $\mathcal{P}^*(x)$ is represented as:

$$\mathcal{P}^*(x) = \beta\mathcal{P}^c(x) + (1-\beta)\left[\gamma\mathcal{P}^t(x) + (1-\gamma)\mathcal{P}^s(x)\right]. \tag{15}$$

IV. Dual Structural Knowledge Interaction

With the structural knowledge, we propose a novel Dual Structural Knowledge Interaction (DSKI) model for domain adaptation, as shown in Fig. 2. For increasing the diversity of structural knowledge of unlabeled target samples, DSKI utilizes two types of structural knowledge to aggregate and interact in weak and strong feature spaces. Specifically, given the target samples $\mathbf{D}_t = \{(x_i^t)\}_{i=1}^{N_t}$, DSKI applies the weak augmentation and strong augmentation strategies to augment the target samples for forming weak and strong feature space, *e.g.,* $\mathbf{D}_1^t = \{(x_{i,1}^t)\}_{i=1}^{N_t}$ and $\mathbf{D}_2^t = \{(x_{i,2}^t)\}_{i=1}^{N_t}$ denote the corresponding weak and strong augmented samples, respectively. The detailed illustrations about the weak augmentation and strong augmentation strategies are shown in the experimental section.

In this work, two independent networks are optimized with $\mathbf{D}_1^t$ and $\mathbf{D}_2^t$ for generating the weak and strong augmented feature spaces. Based on the augmented samples $\mathbf{D}_j^t (j \in \{1,2\})$, DSKI applies the corresponding feature extractor $\mathcal{F}_j$ to extract the target features $\mathbf{F}_j^t = \{\mathbf{f}_{i,j}^t\}_{i=1}^{N_t} (j \in \{1,2\})$. After that, we apply the classifier $\mathcal{C}_j$ to predict the class probabilities $\mathbf{P}_j^t = \{\mathbf{p}_{i,j}^t\}_{i=1}^{N_t}$. For domain adaptation, we also generate the source features $\mathbf{F}_j^s = \{\mathbf{f}_{i,j}^s\}_{i=1}^{N_s}$ and the corresponding class probabilities $\mathbf{P}_j^s = \{\mathbf{p}_{i,j}^s\}_{i=1}^{N_s}$ for the source samples $\mathbf{D}_s = \{(x_i^s)\}_{i=1}^{N_s}$. With the target features $\mathbf{F}_j^t$ and class probabilities $\mathbf{P}_j^t (j \in \{1,2\})$, we can generate and update the target cluster centers $\mathbf{c}$ with (7). Specifically, during training with mini-batch strategy, we employ memory banks to store the source features, target features, and target centers, *e.g.,* as shown in Fig. 2, the memory bank is utilized for each space. After that, we can generate the final structural probabilities $\mathcal{P}^1(x)$ and $\mathcal{P}^2(x)$ of the target sample x for two spaces. Since the structural probabilities $\mathcal{P}^1(x)$ and $\mathcal{P}^2(x)$ describe the same target sample x, we apply a structural knowledge interaction between two feature spaces for knowledge consistency. For example, each space is optimized by the structural knowledge probability of its peer space. Specifically, the structural probabilities $\mathcal{P}^1(x)$ and $\mathcal{P}^2(x)$ are used to optimize strong and weak feature embeddings, respectively. Therefore, the objective of the weak augmented feature space is:

$$\mathcal{L}_{t1} = -\sum_{x_{i,1}^t \in \mathbf{D}_1^t} \tilde{s}_{i,2}^t \tilde{y}_{i,2}^t \log \mathcal{C}_1(\mathcal{F}_1(x_{i,1}^t)), \tag{16}$$

where $\tilde{s}_{i,2}^t$ and $\tilde{y}_{i,2}^t$ denote the maximum probability and its index of the structural probability $\mathcal{P}^2(x_{i,2}^t)$. Similarly, the objective of the strong augmented feature space is:

$$\mathcal{L}_{t2} = -\sum_{x_{i,2}^t \in \mathbf{D}_2^t} \tilde{s}_{i,1}^t \tilde{y}_{i,1}^t \log \mathcal{C}_2(\mathcal{F}_2(x_{i,2}^t)). \tag{17}$$

Except for the structural knowledge alignment, we also maximize the mutual information [60], [61] between two types of instance descriptions to conduct instance mutual alignment for promoting the discrimination of target representations. Note that the propose of utilizing mutual information is different from self-supervised methods, *e.g.*, InfoMin [62] resorts to mutual information to minimize information shared between views and maximize task-relevant information for contrastive learning. In order to calculate the mutual constraint, we firstly convert the target probability sets $\mathbf{P}_1^t$ and $\mathbf{P}_2^t$ into the corresponding probability matrices P_1^t and P_2^t. After that, we can obtain the joint probability distribution matrix P with (18),

$$P = {P_1^t}^\top P_2^t. \tag{18}$$

As for each $(x_{i,1}^t, x_{i,2}^t)$ we also have $(x_{i,2}^t, x_{i,1}^t)$ by considering symmetric problems. Specifically, we convert P into symmetric matrices $\hat{P} = (P + P^\top)/2$. Then, the mutual information loss is defined as (19),

$$\mathcal{L}_m = -\mathcal{I}(P_1^t, P_2^t) = -\sum_{i=1}^{N_c}\sum_{j=1}^{N_c} \hat{P}_{i,j} \cdot \ln \frac{\hat{P}_{i,j}}{\hat{P}_i \cdot \hat{P}_j}, \tag{19}$$

where $\mathcal{I}$ depicts mutual information, $\hat{P}_{i,j}$ denotes each element in row i and column j of $\hat{P}$, $\hat{P}_i$ and $\hat{P}_j$ indicate the marginal distributions which are the summed of the rows and columns of $\hat{P}$, respectively.

By combining the knowledge consistency constraint and mutual constraint, the final objective is:

$$\mathcal{L} = \mathcal{L}_{s1} + \mathcal{L}_{s2} + \zeta(\mathcal{L}_{t1} + \mathcal{L}_{t2}) + \eta\mathcal{L}_m, \tag{20}$$

where ζ and η are the trade-off parameters, $L_{s1} = -\sum_{x_i^s \in \mathbf{D}_s} \bar{y}_i^s \log \mathcal{C}_1(\mathcal{F}_1(x_i^s))$ and $L_{s2} = -\sum_{x_i^s \in \mathbf{D}_s} \bar{y}_i^s \log \mathcal{C}_2(\mathcal{F}_2(x_i^s))$ denote the standard source classification losses with label-smoothing regularization [58] for two spaces, respectively, $\bar{y}_i^s = 0.9 * y_i^s + 0.1/N_c$ is the smoothed label as in SHOT [70].

V. Experiments

A. Datasets

We evaluate the proposed method on three standard domain adaptation benchmarks:

1) *Office-Home* [74] is a challenging medium-sized benchmark with 65 object categories, and consists of 4 different domains: Art (**Ar**), Clipart (**Cl**), Product (**Pr**) and Real-World (**Rw**). We conduct twelve transfer tasks for evalutions.

2) *VisDA-2017* [75] is a challenging large-scale banchmark with 12 object categories, and consists of 2 dissimilar domains: Synthetic and Real. We evalulate the proposed method on **Synthetic** $\rightarrow$ **Real** transfer task.
3) *Office-31* [76] is a widely used benchmark with 31 object categoires, and contains 3 distinct domains: Amazon (**A**), Webcam (**W**) and DSLR (**D**). We evalulate the proposed method on six transfer tasks.

B. Implementation Details

We implement our approach in PyTorch using Nvidia TESLA V100 GPU. ResNet-50 [63] is treated as the backbone for Office-31 and Office-Home, and ResNet-101 [63] is treated as the backbone for VisDA-2017. The last FC layers of ResNet-50 and ResNet-101 are replaced with task-specific FC layers. We utilize all labeled source samples and unlabeled target samples during training. The mixup [77] augmentation is applied to the target samples on Office-31 and VisDA-2017. During inference, the outputs of the classifier in each network are treated as the final predictions.

We set $K = 5$ to select the neighbors of the target sample. The mini-batch SGD with momentum of 0.9 and weight decay of $1\mathrm{e}^{-3}$ is used for training. The learning rate strategy is set as same as CDAN [1], *i.e.*, $\psi_p = \psi_0(1 + ap)^{-b}$, where ψ_p is the current learning rate, ψ_0 is the initial learning rate, p is the training progress changing from 0 to 1, $a = 10$, and $b = 0.75$. The initial learning rate and hyperparameters have different settings among different datasets. For Office-Home, we adopt the initial learning rate of $1\mathrm{e}^{-3}$ and $1\mathrm{e}^{-2}$ for the feature extractors and the classifiers, respectively. We set $\beta = 0.3$, $\gamma = 0.7$, $\epsilon = 0.7$, and utilize a linear rampup scheduler from 0 to 0.2 and 1 for ζ and η, respectively. For Office-31, we set $\beta = 0.5$, $\gamma = 0.5$, and $\epsilon = 0.9$. The settings of ζ, η, and initial learning rate are same as Office-Home. For VisDA-2017, we set $\beta = 0.5$, $\gamma = 0.7$, $\epsilon = 0.7$, $\zeta = 0.2$, and $\eta = 0.8$. For the weak space, the learning rate is set as $5\mathrm{e}^{-4}$ and $5\mathrm{e}^{-3}$ for the feature extractor and the classifier. For the strong space, the learning rate is set as $2\mathrm{e}^{-3}$ and $2\mathrm{e}^{-2}$ for the feature extractor and the classifier.

DSKI applies the weak augmentation and strong augmentation strategies for augmenting the target samples to form weak space and strong space, respectively. The weak augmentation is the flip-and-shift augmentation strategy. For strong augmentation, we adopt RandAugment [78] to sequentially apply several label-preserving image transformations randomly sampled from a predefined set of transforms, *e.g.*, image rotation and contrast adjustment. Note that IL2A [79] also proposes dual augmentation framework for class-incremental learning. However, its dual augmentation consists of class augmentation and semantic augmentation in sample space and feature space to synthesize auxiliary classes and mimic the distribution of old classes, respectively, which is different from the proposed method.

C. Comparison With Existing Methods

In this section, we conduct comparisons between the proposed DSKI and classic pseudo-labeling methods in semi-supervised learning (SSL) and existing domain adaptation (DA) methods on Office-Home, VisDA-2017, and Office-31, and summarize the related results in Tables I–III, respectively. Since two independent networks are optimized with weak and strong augmented target samples for generating the weak and strong augmented feature spaces, we denote these two networks as weak and strong networks, respectively. DSKI-W and DSKI-S represent the results in weak network and strong network, respectively.

Table I illustrates the classification results on twelve tasks of Office-Home. Specifically, DSKI-W/DSKI-S outperforms Baseline trained with labeled source domain only with large improvements, *i.e.*, 28.7%/28.8% in average accuracy. Moreover, DSKI-W/DSKI-S also achieves notable improvements compared with semi-supervised methods, *e.g.*, outperforming Fixmatch by 5.2%/5.3% in average accuracy. Furthermore, the proposed method boosts the results upon the state-of-the-art domain adaptation method, *e.g.*, DSKI-W and DSKI-S achieve 2.6% and 2.7% mean improvement than ATDOC-NA, respectively.

Table II depicts the accuracy on twelve classes of VisDA-2017. Particularly, compared with Baseline, DSKI-W/DSKI-S obtains the improvements of 35.3%/35.3% in average accuracy. Moreover, compared with classic pseudo-labeling methods in semi-supervised learning and existing domain adaptation, DSKI-W/DSKI-S achieves the best-performing results, *i.e.*, 87.7%/87.7% in average accuracy.

Table III lists the results on six tasks of Office-31. We observe that DSKI-W/DSKI-S outperforms Baseline by 14.7%/14.7% in average accuracy. Moreover, DSKI-W/DSKI-S gains the best results in challenging D $\rightarrow$ A and W $\rightarrow$ A tasks. Note that DSKI-W/DSKI-S obtains a slight improvement (*e.g.* 0.2%/0.2%) compared with the state-of-the-art domain adaptation method CAN on Office-31, because each domain in Office-31 contains fewer images and smaller inter-domain differences than other dataset.

Although we apply two independent feature spaces for target feature learning, we also observe that the weak and strong networks achieve the similar performance on all three benchmarks, *e.g.,* 74.8% *vs.* 74.9%, 87.7% *vs.* 87.7%, and 90.8% *vs.* 90.8% for Office-Home, VisDA-2017, and Office-31, respectively. The reason is that we apply a knowledge consistency constraint to conduct structural knowledge interaction between two feature spaces, leading to two feature embedding producing similar target descriptions finally.

D. Ablation Study

Knowledge Aggregation: The structural knowledge consists of cluster-based knowledge (C), target locality-based knowledge (TL), and source locality-based knowledge (SL). We conduct several experiments to evaluate the effectiveness of the proposed components. As shown in Table IV, using the cluster-based knowledge(C) obtains a higher performance than the pesudo-labeling method using implicitly source cluster centers, *e.g.,* 89.5%/89.5% *vs.* 87.3%/87.2% in weak/strong network. Moreover, utilizing target locality-based knowledge (TL) obtains a higher performance than using implicitly source cluster centers, *e.g.,* TL-W/TL-S improves the mean accuracy of pseudo-labeling method from 87.3%/87.2% to 90.1%/90.1%. Furthermore, SL-W and SL-S employing the source locality-based knowledge obtain 0.7% and 0.8% improvement compared

TABLE I
CLASSIFICATION ACCURACY (%) ON OFFICE-HOME

	Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
Baseline	ResNet-50 [63]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
SSL	Mean-teacher [64]	47.0	66.9	75.1	57.7	64.4	68.2	55.3	43.3	75.1	67.4	49.3	79.3	62.4
	Mixmatch [65]	53.6	74.6	79.1	65.6	72.8	74.6	62.2	47.9	79.3	73.0	58.1	83.7	68.7
	Fixmatch [12]	51.9	71.9	78.7	65.1	73.3	73.1	66.7	56.5	79.3	74.2	60.3	84.7	69.6
DA	DANN [5]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
	DAN [66]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
	CDAN+E [41]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
	SAFN [67]	52.0	71.7	76.3	64.2	69.9	71.9	63.7	51.4	77.1	70.9	57.1	81.5	67.3
	MCC [68]	56.3	77.3	80.3	67.0	77.1	77.0	66.2	55.1	81.2	73.5	57.4	84.1	71.0
	BNM [69]	56.7	77.5	81.0	67.3	76.3	77.1	65.3	55.1	82.0	73.6	57.0	84.3	71.1
	SHOT [70]	57.1	78.1	81.5	68.0	78.2	78.1	67.4	54.9	82.2	73.3	58.8	84.3	71.8
	HDAN [71]	56.8	75.2	79.8	65.1	73.9	75.2	66.3	56.7	81.8	75.4	59.7	84.7	70.9
	ATDOC-NC [72]	54.4	77.6	80.8	66.5	75.6	75.8	65.9	51.9	81.1	72.7	57.0	83.5	70.2
	ATDOC-NA [72]	58.3	78.8	82.3	69.4	78.2	78.2	67.1	56.0	82.7	72.0	58.2	85.5	72.2
	FGDA [73]	52.3	77.0	78.2	64.6	75.5	73.7	64.0	49.5	80.7	70.1	52.3	81.6	68.3
Ours	DSKI-W	61.6	**80.2**	82.4	72.0	81.2	80.6	**70.3**	61.4	82.9	**76.6**	**62.4**	86.2	74.8
	DSKI-S	**61.8**	**80.2**	**82.5**	**72.2**	**81.3**	**80.7**	**70.3**	**61.7**	**83.0**	76.4	**62.4**	**86.4**	**74.9**

The best performance is highlighted in bold.

TABLE II
CLASSIFICATION ACCURACY (%) ON VISDA-2017

	Method	plane	bcycle	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg.
Baseline	ResNet-101 [78]	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
SSL	Mean-teacher [64]	71.3	52.5	80.2	48.7	77.6	38.2	81.5	54.2	74.4	51.7	76.6	27.7	61.2
	Mixmatch [65]	83.8	67.0	91.2	68.4	94.1	70.6	92.5	80.5	85.7	80.5	81.6	17.3	76.1
	Fixmatch [12]	95.6	67.2	77.3	58.3	94.9	0.1	**93.3**	76.6	90.9	94.0	89.7	18.3	71.3
DA	DANN [5]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
	DAN [66]	87.1	63.0	76.5	42.0	90.3	42.9	85.9	53.1	49.7	36.3	85.8	20.7	61.1
	CDAN [41]	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.9
	SAFN [67]	93.6	61.3	84.1	70.6	94.1	79.0	91.8	79.6	89.9	55.6	89.0	24.4	76.1
	MCC [68]	92.2	82.9	76.8	66.6	90.9	78.5	87.9	73.8	90.1	76.1	87.1	41.0	78.7
	BNM [69]	91.1	69.0	76.7	64.3	89.8	61.2	90.8	74.8	90.9	66.6	88.1	46.1	75.8
	CAN [28]	97.0	87.2	82.5	74.3	**97.8**	**96.2**	90.8	80.7	**96.6**	**96.3**	87.5	**59.9**	87.2
	SHOT [70]	94.3	88.5	80.1	57.3	93.1	94.9	80.7	80.3	91.5	89.1	86.3	58.2	82.9
	BCDM [46]	95.1	87.6	81.2	73.2	92.7	95.4	86.9	82.5	95.1	84.8	88.1	39.5	83.4
	ATDOC-NC [72]	91.1	60.1	78.4	72.2	88.1	97.6	86.9	55.9	79.2	64.9	88.4	31.9	74.6
	ATDOC-NA [72]	93.7	83.0	76.9	58.7	89.7	95.1	84.4	71.4	89.4	80.0	86.7	55.1	80.3
Ours	DSKI-W	96.8	**89.7**	89.0	**80.2**	97.1	96.1	91.6	83.4	95.5	94.5	91.1	47.6	**87.7**
	DSKI-S	**97.2**	88.4	**89.4**	79.6	96.9	95.8	91.8	**84.0**	95.8	95.2	**91.2**	47.5	**87.7**

The best performance is highlighted in bold.

TABLE III
CLASSIFICATION ACCURACY (%) ON OFFICE-31

	Method	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
Baseline	ResNet-50 [63]	68.9	68.4	62.5	96.7	60.7	99.3	76.1
SSL	Mean-teacher [64]	80.7	75.7	65.0	96.7	63.6	99.8	80.3
	Mixmatch [65]	87.6	88.4	65.3	99.2	67.3	99.6	84.6
	Fixmatch [12]	88.6	85.2	66.7	98.1	62.5	**100.0**	83.5
DA	DANN [5]	79.7	82.0	68.2	96.9	67.4	99.1	82.2
	DAN [66]	78.6	80.5	63.6	97.1	62.8	99.6	80.4
	CDAN+E [41]	92.9	94.1	71.0	98.6	69.3	**100.0**	87.7
	SAFN+ENT [67]	90.7	90.1	73.0	98.6	70.2	99.8	87.1
	MCC [68]	92.1	94.0	74.9	98.5	75.3	**100.0**	89.1
	BNM [69]	92.2	94.0	74.9	98.5	75.3	**100.0**	89.2
	CAN [28]	95.0	94.5	**78.0**	99.1	77.0	99.8	90.6
	SHOT [70]	94.0	90.1	74.7	98.4	74.3	99.9	88.6
	BCDM [46]	93.8	**95.4**	73.1	98.6	73.0	**100.0**	89.0
	ATDOC-NC [72]	95.2	91.6	74.6	99.1	74.7	**100.0**	89.2
	ATDOC-NA [72]	94.4	94.3	75.6	98.9	75.2	99.6	89.7
Ours	DSKI-W	**96.0**	94.6	77.7	**99.2**	77.3	**100.0**	**90.8**
	DSKI-S	**96.0**	94.6	77.6	99.1	**77.6**	**100.0**	**90.8**

The best performance is highlighted in bold.

TABLE IV
ABLATION STUDY OF KNOWLEDGE AGGREGATION ABOUT ACCURACY ON OFFICE-31

Method	C	TL	SL	A→D	A→W	D→A	D→W	W→A	W→D	Avg
Pseudo-labeling-W				91.4	90.9	72.2	98.7	70.5	**100.0**	87.3
C-W	✓			95.4	93.1	74.7	**99.2**	74.3	**100.0**	89.5
TL-W		✓		95.6	94.5	74.9	**99.2**	76.4	**100.0**	90.1
SL-W			✓	92.4	91.3	74.2	98.4	71.7	**100.0**	88.0
DSKI-W	✓	✓	✓	**96.0**	**94.6**	**77.7**	**99.2**	**77.3**	**100.0**	**90.8**
Pseudo-labeling-S				91.6	90.9	72.0	98.5	70.5	**100.0**	87.2
C-S	✓			95.4	93.2	74.7	**99.1**	74.6	**100.0**	89.5
TL-S		✓		95.6	94.5	75.2	**99.1**	76.4	**100.0**	90.1
SL-S			✓	92.4	91.4	74.4	98.2	71.8	**100.0**	88.0
DSKI-S	✓	✓	✓	**96.0**	**94.6**	**77.6**	**99.1**	**77.6**	**100.0**	**90.8**

TABLE V
THE ACCURACY OF PSEUDO-LABELS ABOUT VARIOUS PSEUDO-LABELING METHODS ON OFFICE-31

Method	A→D	A→W	D→A	D→W	W→A	W→D	Avg
Mean-teacher [64]	76.9	72.6	56.9	95.7	56.9	99.6	76.4
Mixmatch [65]	88.5	88.5	69.3	**99.0**	64.9	99.6	85.0
Fixmatch [12]	87.4	86.2	63.2	97.5	62.0	**100.0**	82.7
SHOT [70]	91.4	87.4	71.3	98.1	73.8	99.6	86.9
ATDOC-NC [72]	95.3	91.1	73.0	98.9	74.3	99.8	88.7
ATDOC-NA [72]	93.4	92.9	75.0	97.9	**75.3**	98.7	88.9
DSKI	**95.7**	**93.7**	**75.1**	98.7	74.9	**100.0**	**89.7**

with using implicitly source cluster centers, respectively. Finally, jointly considering three knowledge terms, the mean accuracies of DSKI-W and DSKI-S reach 90.8% and 90.8%, respectively. The above analyses demonstrate the effectiveness of each term and the necessity of integrating the three terms.

The Accuracy of Pseudo-labels: To compare the accuracy of pseudo-labels about various pseudo-labeling methods, we present the accuracy of pseudo-labels among the proposed DSKI and classic pseudo-labeling methods in semi-supervised learning (Mean-teacher, Mixmatch, and Fixmatch) as well as in domain adaptation (SHOT, ATDOC-NC, and ATDOC-NA) on Office-31, as shown in Table V. The accuracies of pseudo-labels about pseudo-labeling methods are obtained by re-implementing the released codes. For fair comparison, the proposed DSKI is implemented by only considering knowledge aggregation and ignoring knowledge consistency constraint and mutual constraint. We observe the pseudo-labeling methods in semi-supervised learning achieve inferior performances in domain adaptation task, *i.e.*, average accuracy 82.7% of Fixmatch, since these methods designed for singe domain scenario cannot handle cross domain scenario. Moreover, the pseudo-labeling methods in domain adaptation obtain higher results than the pseudo-labeling methods in semi-supervised learning, but their results are all lower than the proposed DSKI, *i.e.*, average accuracy 88.9% of ATDOC-NA *vs.* average accuracy 89.7% of DSKI. This is because these methods only consider cluster-based knowledge or target locality-based knowledge, and ignore source locality-based knowledge. By jointly considering cluster-based knowledge, source locality-based knowledge, and target locality-based knowledge, DSKI achieves the best results about average accuracy of pseudo-labels, demonstrating the effectiveness of the proposed DSKI.

We further compare the accuracy of pseudo-labels among DSKI and the methods using various knowledge in DSKI, and present the results in Table VI, where Pseudo-labeling represents obtaining pseudo-labels with implicit source cluster centers, C depicts utilizing cluster-based knowledge for pseudo-labels, TL and SL indicate considering target locality-based knowledge and source locality-based knowledge, respectively. We observe that the average accuracies (89.1%/88.5%, 89.9%/88.8%, and 87.7%/87.1%) of pseudo-labels obtained by C, TL, and SL are higher than the average accuracy (86.9%/86.0%) of pseudo-labels obtained by pseudo-labeling in weak/strong space, illustrating the effectiveness of target cluster centers, target locality-based knowledge, and source locality-based knowledge. Note that SL acquires lower average accuracy of pseudo-labels than TL and C, since the distributions of source domain and target domain are different. Combining cluster-based knowledge, source locality-based knowledge, and target locality-based knowledge, DSKI achieves the best results about average accuracy (90.6%/89.4%) of pseudo-labels in weak/strong space, illustrating the rationality of knowledge aggregation.

Effect of Knowledge Consistency Constraint $\mathcal{L}_{t*}$: After obtaining the structural knowledge in two independent networks, the knowledge consistency constraint $\mathcal{L}_{t*}$ is used for structural knowledge interaction between two feature spaces. As shown in Table VII, the baseline model (SKA) represents the model only considers knowledge aggregation and ignores knowledge consistency constraint $\mathcal{L}_{t*}$ and mutual constraint $\mathcal{L}_m$. Specifically, SKA adopts the obtained pseudo-label to train its own space, rather than its peer space. Moreover, SKIA represents utilizing the knowledge consistency constraint $\mathcal{L}_{t1}$ and $\mathcal{L}_{t2}$ to optimize each space embedding in SKA. From Table VII we can see that, SKA obtains the mean accuracy of 73.3% and 72.2% in weak and strong networks, respectively. After adding the knowledge consistency constraint to SKA, SKIA obtains the mean accuracies of 73.6% and 73.9% in weak and strong networks, respectively. The improvements demonstrate that using the unsupervised

TABLE VI
THE ACCURACY OF PSEUDO-LABELS ABOUT METHODS USING VARIOUS KNOWLEDGE IN DSKI ON OFFICE-31

Method	C	TL	SL	A→D	A→W	D→A	D→W	W→A	W→D	Avg
Pseudo-labeling-W				91.2	90.7	70.9	98.6	69.7	**100.0**	86.9
C-W	✓			95.5	92.9	72.9	**98.9**	74.3	99.8	89.1
TL-W		✓		95.7	94.4	74.9	**98.9**	75.5	**100.0**	89.9
SL-W			✓	92.3	91.4	73.3	98.1	71.2	**100.0**	87.7
DSKI-W	✓	✓	✓	**96.2**	**94.6**	**77.0**	**98.9**	**76.8**	**100.0**	**90.6**
Pseudo-labeling-S				90.8	90.0	69.1	98.0	68.2	**99.8**	86.0
C-S	✓			95.5	92.6	72.4	97.6	72.8	**99.8**	88.5
TL-S		✓		95.5	94.2	71.3	**98.7**	**74.1**	99.1	88.8
SL-S			✓	92.3	90.8	72.0	97.7	70.1	99.6	87.1
DSKI-S	✓	✓	✓	**96.2**	**93.9**	**75.0**	97.9	**74.1**	99.1	**89.4**

TABLE VII
ANALYSIS OF INTERACTION IN OFFICE-HOME

Methods	Weak network	Strong network
SKA	73.3	72.2
SKIA	73.6	73.9
SKIM	74.7	74.2
DSKS	73.6	73.8
DSKI	**74.8**	**74.9**

TABLE VIII
ANALYSIS OF AUGMENTATION IN OFFICE-HOME

Methods	Network 1	Network 2
SKBS	73.3	73.3
SKBW	74.5	74.5
SKBH	74.3	74.3
DSKI	**74.8**	**74.9**

knowledge consistency constraint can exchange knowledge between two spaces for inferring robust target representations.

Effect of Instance Mutual Constraint $\mathcal{L}_m$: Besides the knowledge consistency constraint, the instance mutual constraint $\mathcal{L}_m$ is also applied for representation learning. As shown in Table VII, SKIM using the mutual constraint $\mathcal{L}_m$ obtains the averaged 1.4% and 2.0% improvements upon the baseline model (SKA) in two networks. Especially, the final model both considering the knowledge consistency contraint and instance mutual constraint achieves the best mean accuracy of 74.8% and 74.9%, outperforming DSKS which replaces the mutual information constraint with consistent constraint [12]. The better performance shows that the instance mutual constraint is complementary to the knowledge consistency constraint for boosting the target representation learning.

Effect of Dual Space: To evaluate the effect of space embedding for domain adaptation, we conduct several experiments with different augmentation strategies to construct two spaces: both weak (SKBW), both strong (SKBS), and both hybrid augmentations (SKBH), respectively. Note that the hybrid augmentation applies the weak augmentation and strong augmentation to augment the unlabeled target samples. As shown in Table VIII, the weak augmentation strategy is more effective than strong augmentation, *e.g.,* SKBW obtains a higher performance of 74.5% than 73.3% of SKBS. Furthermore, SKBH achieves the medium performance between SKBS and SKBW, *e.g.,* 74.3% of SKBH *vs.* 73.3% of SKBS, and 74.3% of SKBH *vs.* 74.5% of SKBW. The above results demonstrate the intensity of data augmentation during training affects the final performance. We also observe that the proposed DSKI consisting of weak and strong augmentations obtains the best performance of 74.8% and 74.9% in two spaces, proving the necessity and rationality of the interaction between two different augmented spaces.

Hyperparameter Analysis: We conduct the hyperparameter analyses, and summarize the related results in Fig. 3. Fig. 3(a) shows that a proper neighboring size $K = 5$ is important for exploiting structural knowledge. The reason is that small K leads to limited locality-based knowledge information, and large K brings much irrelated label information to structural knowledge. Fig. 3(b) illustrates that the trade-off parameter $\beta = 0.5$ between cluster-based knowledge and locality-based knowledge results in the best performance. The results show cluster-based knowledge and locality-based knowledge contribute to structural knowledge equally. Fig. 3(c) indicates that the trade-off parameter $\gamma = 0.5$ between the target locality-based knowledge and source locality-based knowledge achieves the highest accuracy. We find that giving more attention to locality-based knowledge would harm the performance due to the domain gap between source and target domains. Fig. 3(d) depicts that the weight $\zeta = 0.2$ for knowledge alignment loss outperforms other setting. When ζ increases, the wrong information contained in probability predictions is amplified. Furthermore, when ζ decreases, the useful structural knowledge is not fully utilized. Fig. 3(e) represents that proper weight $\eta = 1$ is important to balance mutual information losses.

Analysis of Manifold Assumption: We denote the average rate of selected neighborhood set having same label with samples as homophily score, and present the homophily score of selected target neighborhood set and selected source neighborhood set in Fig. 4. We observe that as the neighboring size K decreases, the homophily score of selected target neighborhood set increases, demonstrating the manifold assumption that the higher similarity, the higher confidence with same labels. Moreover, the homophily score of selected source neighborhood set almost satisfies the manifold assumption as K changes, except $K = 3$. This is because there is domain gap between source domain and target domain, the nearby source samples of target sample may have inconsistent labels. The above results illustrate the rationality of the manifold assumption in DSTK.

Model Analysis: We conduct some ablation studies for model analysis, and show the results in Table IX. Firstly, the structural knowledge considers the source neighbors and target neighbors of the target sample to construct locality-based knowledge. Note that the target neighbors do not contain the target sample itself

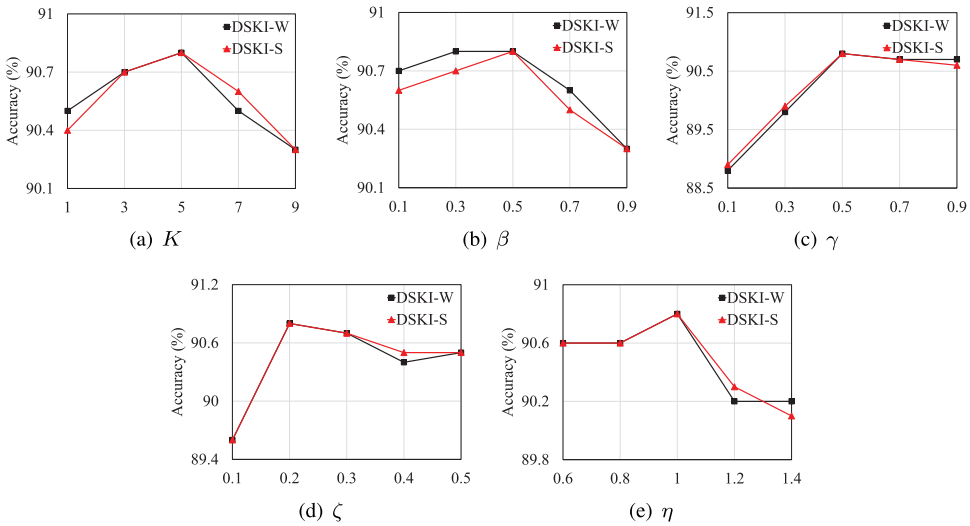


Fig. 3. Plot of various hyperparameter analyses on Office-31. (a) Varying the neighboring size K, the number of selected neighbors about each sample for calculating locality-based knowledge. (b) Varying the trade-off parameter β between cluster-based knowledge and locality-based knowledge for obtaining structural knowledge. (c) Varying the trade-off parameter γ between the target locality-based knowledge and source locality-based knowledge for gaining locality-based knowledge. (d) Varying the weight ζ for structural knowledge alignment loss. (e) Varying the weight η for mutual information loss.

TABLE IX
ANALYSIS ABOUT WHETHER UTILIZING LABEL INFORMATION FOR EACH SAMPLE ON OFFICE-31

Method	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
DSKI-W w/ T	95.0	94.5	77.6	**99.2**	**77.4**	**100.0**	90.6
DSKI-W	**96.0**	**94.6**	**77.7**	**99.2**	77.3	**100.0**	**90.8**
DSKI-S w/ T	95.0	94.5	**77.8**	**99.1**	77.4	**100.0**	90.6
DSKI-S	**96.0**	**94.6**	77.6	**99.1**	**77.6**	**100.0**	**90.8**

The best performance is highlighted in bold.

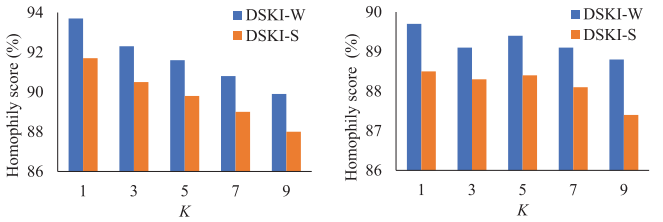


Fig. 4. The homophily scores of selected target neighborhood set and selected source neighborhood set on Office-31.

due to its label might be incorrect. We thus perform a comparison between with and without considering the target sample itself for the locality-based knowledge, where "DSKI w/ T" and DSKI represent the models with and without considering the target sample itself, respectively. As shown in Table IX, "DSKI-W w/T" and "DSKI-S w/T" obtain the mean accuracy of 90.6% and 90.6% in weak and strong networks, which are lower than 90.8% of DSKI-W and 90.8% of DSKI-S. This poor performance of "DSKI w/ T" indicates that the pseudo-labels of the target images contain much wrong label information. Therefore, the locality-based knowledge considering its own label of the target sample is not conducive to generate robust pseudo-labels.

Secondly, the cluster-based knowledge constructs a structure to aggregate all target cluster center information, which is different from the traditional methods by treating the label of the most similar cluster center as a pseudo label. To evaluate the effectiveness of the cluster-based knowledge, we thus perform a comparison between DSKI and "DSKI w/ C" which constructs the cluster-based knowledge with the most similar target cluster center. As shown in Table X, DSKI-W and DSKI-S obtain 1.2% and 1.2% improvements upon the "DSKI-W w/ C" and "DSKI-S w/ C," which demonstrates the necessity and effectiveness to apply the structure to fuse all cluster centers.

Finally, we adopt hard probability as center probability $\mathbf{p}_i^c$ for target centers. Note that for the center probability, there are two situations: soft probability and hard probability. We compare the results between soft probability and hard probability of center probability $\mathbf{p}_i^c$ for target centers, as shown in Table XI, where DSKI w/Soft denotes utilizing the soft probability obtained by feeding the target center feature $\mathbf{c}_i$ into classifier $\mathcal{C}$, and DSKI adopts hard probability using a one-hot vector for describing the center belonging to each class. We observe that DSKI outperforms DSKI w/Soft by 0.5% and 0.4% in weak space and strong space, respectively, indicating hard probability is more effective than soft probability for target centers.

Precision, Recall, and F1-score: For achieving more comprehensive insights, we further compare the precision, recall, and F1-score of the proposed method with several classic domain adaptation methods (DAN, DANN, MCC, BNM, SHOT, ATDOC-NC and ATDOC-NA), as shown in Table XII. The precision, recall, and F1-score of domain adaptation methods are

TABLE X
ANALYSIS ABOUT THE WAY TO OBTAIN THE CLUSTER-BASED KNOWLEDGE ON OFFICE-31

Method	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
DSKI-W w/ C	95.4	91.9	75.9	**99.2**	75.2	**100.0**	89.6
DSKI-W	**96.0**	**94.6**	**77.7**	**99.2**	**77.3**	**100.0**	**90.8**
DSKI-S w/ C	95.6	91.8	75.9	**99.1**	75.4	**100.0**	89.6
DSKI-S	**96.0**	**94.6**	**77.6**	**99.1**	**77.6**	**100.0**	**90.8**

The best performance is highlighted in bold.

TABLE XI
ANALYSIS ABOUT PROBABILITY OF TARGET CENTERS ON OFFICE-31

Method	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
DSKI-W w/ Soft	95.4	94.2	76.4	99.1	76.8	**100.0**	90.3
DSKI-W	**96.0**	**94.6**	**77.7**	**99.2**	**77.3**	**100.0**	**90.8**
DSKI-S w/ Soft	95.4	94.2	76.9	**99.1**	76.9	**100.0**	90.4
DSKI-S	**96.0**	**94.6**	**77.6**	**99.1**	**77.6**	**100.0**	**90.8**

The best performance is highlighted in bold.

TABLE XII
THE COMPARSION ABOUT PRECISION, RECALL, AND F1-SCORE ON OFFICE-31

Method	Precision	Recall	F1-score
DAN [5]	0.831	0.829	0.818
DANN [66]	0.811	0.802	0.790
MCC [68]	0.885	0.881	0.873
BNM [69]	0.898	0.899	0.892
SHOT [70]	0.869	0.873	0.864
ATDOC-NC [72]	0.905	0.892	0.890
ATDOC-NA [72]	0.893	0.892	0.886
DSKI-W	0.912	**0.909**	**0.905**
DSKI-S	**0.913**	0.908	**0.905**

obtained by re-implementing the released codes. We observe that some methods achieve high precision, but obtain relative low recall, *e.g.*, the precision (0.905) of ATDOC-NC is higher than the precision (0.898) of BNM, but the recall (0.892) of ATDOC-NC is lower than the recall (0.899) of BNM. However, compared with other methods, the proposed DSKI achieves the best-performing precision and recall, *i.e.*, 0.912/0.913 of precision and 0.909/0.908 of recall in weak/strong space. Moreover, the F1-score of the proposed DSKI is also highest, *i.e.*, 0.905/0.905 of F1-score in weak/strong space. The above results demonstrate the effectiveness of the proposed DSKI.

Visualization: Fig. 5 shows the target feature visualization with t-SNE [2] of **D** → **A** in Office-31. As shown in Fig. 5(a), merely using the source samples to optimize the feature extractor leads to that the target features are misaligned with source features. With the help of pseudo-labels, the target features are slightly aligned with the source features, but the target features are not discriminative, as shown in Fig. 5(b). After aggregating and interacting the knowledge in the source and target domains, the target features and source features are well aligned, and the target features are discriminative, shown in Fig. 5(c) and (d).

Time and Space Complexity: Since reducing the time/space complexity is not our main purpose and DSKI adopts two networks for dual structural knowledge interaction, the time/space complexity is inevitable higher than the methods with single network, such as SHOT. However, the performance of DSKI is much higher than SHOT, *e.g.*, improving the accuracy from 74.7% to 77.7% in **D** → **A** task on Office-31. Moreover, the runtime and extra space size of DSKI are acceptable shown in

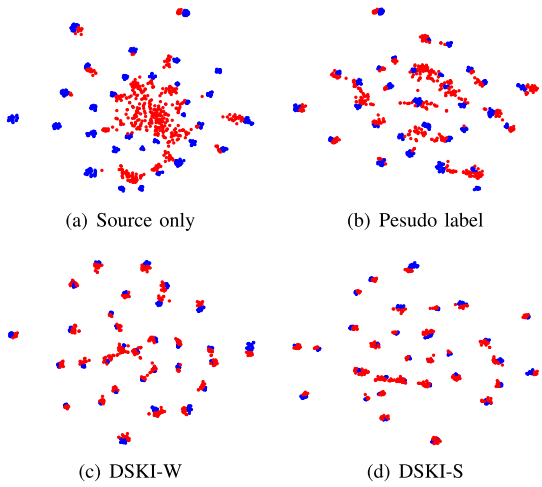


Fig. 5. Feature visualization of the **D** → **A** tasks in Office-31. Blue and Red dots represent the source features and target features, respectively.

TABLE XIII
TIME AND SPACE COMPLEXITY IN **D** → **A** TASK ON OFFICE-31

Time complexity (Runtime)	Space complexity (Extra space size)		
Per epoch	C	SL	TL
79.3s	0.03M	0.27M	1.54M

Table XIII, *i.e.*, the runtime is 79.3 s per epoch, the extra space sizes for cluster-based knowledge (C), source locality-based knowledge (SL), and target locality-based knowledge (TL) in memory bank are 0.03 M, 0.27 M, and 1.54 M, respectively.

VI. CONCLUSION

In this work, we introduce a novel structural knowledge for pseudo-labeling the target samples, and further propose a novel Dual Structural Knowledge Interaction (DSKI) framework for domain adaptation. The structural knowledge is proposed by aggregating the cluster-based knowledge, source locality-based knowledge and target locality-based knowledge to obtain structural knowledge. Moreover, DSKI adopts two different spaces to interact the structural knowledge for target representation learning. Furthermore, we also maximize the mutual information between the weak and strong target descriptions to generate the

aligned and discriminative features. The evaluations of three benchmarks verify the effectiveness of the proposed method. In the future, we will utilize the graph convolutional network (GCN) to efficiently aggregate the knowledge descriptions.

REFERENCES

[1] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.

[2] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1410–1417.

[3] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 136–144.

[4] J. Li et al., "Maximum density divergence for domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3918–3930, Nov. 2021.

[5] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, pp. 2096–2030, 2016.

[6] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5419–5428.

[7] Y. Wu, D. Inkpen, and A. El-Roby, "Dual mixup regularized learning for adversarial domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*2020, pp. 540–555.

[8] W. Zhang, D. Xu, W. Ouyang, and W. Li, "Self-paced collaborative and adversarial network for unsupervised domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2047–2061, Jun. 2021.

[9] T. Chen et al., "Enhanced feature alignment for unsupervised domain adaptation of semantic segmentation," *IEEE Trans. Multimedia*, vol. 24, pp. 1042–1054, 2022.

[10] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *Proc. Int. Joint Conf. Neural Netw.*, 2020, pp. 1–8.

[11] W. Shi, Y. Gong, C. Ding, Z. M. Tao, and N. Zheng, "Transductive semi-supervised deep learning using min-max features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 299–315.

[12] K. Sohn et al., "Fixmatch: Simplifying Semi-Supervised Learning With Consistency and Confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., 2020, pp. 596–608.

[13] W.-G. Chang, T. You, S. Seo, S. Kwak, and B. Han, "Domain-specific batch normalization for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7354–7362.

[14] Z. Deng, Y. Luo, and J. Zhu, "Cluster alignment with a teacher for unsupervised domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9944–9953.

[15] C. Chen et al., "Progressive feature alignment for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 627–636.

[16] Y. Zuo, H. Yao, L. Zhuang, and C. Xu, "Seek common ground while reserving differences: A model-agnostic module for noisy domain adaptation," *IEEE Trans. Multimedia*, vol. 24, pp. 1020–1030, 2022.

[17] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "End-to-end memory networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2440–2448.

[18] J. Na, H. Jung, H. J. Chang, and W. Hwang, "FixBi: Bridging domain spaces for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1094–1103.

[19] Y. Zhang, T. Liu, M. Long, and M. Jordan, "Bridging theory and algorithm for domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7404–7413.

[20] S. Chen, M. Harandi, X. Jin, and X. Yang, "Domain adaptation by joint distribution invariant projections," *IEEE Trans. Image Process.*, vol. 29, pp. 8264–8277, 2020.

[21] M. Mancini, L. Porzi, S. R. Bulo, B. Caputo, and E. Ricci, "Inferring latent domains for unsupervised deep domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 485–498, Feb. 2021.

[22] J. Li et al., "Maximum density divergence for domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3918–3930, Nov. 2021.

[23] Y. Lu et al., "Discriminative invariant alignment for unsupervised domain adaptation," *IEEE Trans. Multimedia*, vol. 24, pp. 1871–1882, 2022.

[24] H. Yan et al., "Weighted and class-specific maximum mean discrepancy for unsupervised domain adaptation," *IEEE Trans. Multimedia*, vol. 22, pp. 2420–2433, 2020.

[25] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *CoRR*, vol. abs/1412.3474, 2014. [Online]. Available: http://arxiv.org/abs/1412.3474

[26] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis*. 2016, pp. 443–450.

[27] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, "Central moment discrepancy (CMD) for domain-invariant representation learning," in *Proc. Int. Conf. Learn. Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=SkB-_mcel

[28] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4893–4902.

[29] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1853–1865, Sep. 2017.

[30] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Proc. 32nd AAAI Conf. Artif. Intell. 13th Innov. Appl. Artif. Intell. Conf. 8th AAAI Symp. Educ. Adv. Artif. Intell.*, 2018, pp. 4058–4065.

[31] J. Lee and M. Raginsky, "Minimax statistical learning with wasserstein distances," in *Proc. Adv. Neural Inf. Process. Syst.*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, vol. 31, pp. 2692–2701. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/ea8fcd92d59581717e06eb187f10666d-Paper.pdf

[32] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10285–10295.

[33] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy, "Joint distribution optimal transportation for domain adaptation," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3733–3742.

[34] C. Villani, *Optimal Transport: Old and New*, vol. 338. Berlin, Germany: Springer, 2009.

[35] J. Sun, Z. Wang, W. Wang, H. Li, and F. Sun, "Domain adaptation with geometrical preservation and distribution alignment," *Neurocomputing*, vol. 454, pp. 152–167, 2021.

[36] I. Goodfellow et al., "Generative adversarial networks," *J. Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[37] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[38] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Adversarial dropout regularization," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=HJIoJWZCZ

[39] Y.-H. Tsai et al., "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7472–7481.

[40] X. Ma, T. Zhang, and C. Xu, "Deep multi-modality adversarial networks for unsupervised domain adaptation," *IEEE Trans. Multimedia*, vol. 21, pp. 2419–2431, 2019.

[41] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1645–1655.

[42] S. Cui et al., "Gradually vanishing bridge for adversarial domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12455–12464.

[43] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3723–3732.

[44] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2507–2516.

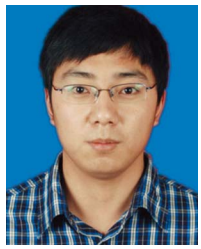
[45] Z. Lu et al., "Stochastic classifiers for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9111–9120.

[46] S. Li et al., "Bi-classifier determinacy maximization for unsupervised domain adaptation," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 8455–8464.

[47] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2200–2207.

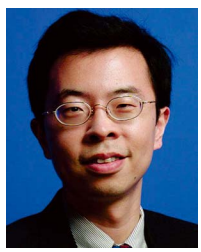
[48] J. Wang et al., "Visual domain adaptation with manifold embedded distribution alignment," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 402–410.
[49] H. Tang, K. Chen, and K. Jia, "Unsupervised domain adaptation via structurally regularized deep clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8725–8735.
[50] J. Liang, D. Hu, and J. Feng, "Domain adaptation with auxiliary target domain-oriented classifier," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16632–16642.
[51] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *Proc. 32nd AAAI Conf. Artif. Intell. 13th Innov. Appl. Artif. Intell. Conf. 8th AAAI Symp. Educ. Adv. Artif. Intell.*, 2018, pp. 3934–3941.
[52] S. Yang et al., "Exploiting the intrinsic neighborhood structure for source-free domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 29393–29405.
[53] K. Saito et al., "Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9184–9193.
[54] N. Ding et al., "Source-free domain adaptation via distribution estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7212–7222.
[55] J. Wu et al., "Clustering and dynamic sampling based unsupervised domain adaptation for person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2019, pp. 886–891.
[56] X. Ma, T. Zhang, and C. Xu, "GCAN: Graph convolutional adversarial network for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8266–8276.
[57] M. Mancini, S. R. Bulo, B. Caputo, and E. Ricci, "AdaGraph: Unifying predictive and continuous domain adaptation through graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6568–6577.
[58] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.
[59] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[60] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9865–9874.
[61] T. Han, J. Gao, Y. Yuan, and Q. Wang, "Unsupervised semantic aggregation and deformable template matching for semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 9972–9982.
[62] Y. Tian et al., "What makes for good views for contrastive learning?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6827–6839.
[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[64] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.
[65] D. Berthelot et al., "MixMatch: A holistic approach to semi-supervised learning," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 5049–5059.
[66] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.
[67] R. Xu, G. Li, J. Yang, and L. Lin, "Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1426–1435.
[68] Y. Jin, X. Wang, M. Long, and J. Wang, "Minimum class confusion for versatile domain adaptation," in *Proc. Eur. Conf. Comput. Vis.* 2020, pp. 464–480.
[69] S. Cui et al., "Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3941–3950.
[70] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6028–6039.
[71] S. Cui, X. Jin, S. Wang, Y. He, and Q. Huang, "Heuristic domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., 2020, pp. 7571–7583.
[72] J. Liang, D. Hu, and J. Feng, "Domain adaptation with auxiliary target domain-oriented classifier," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16632–16642.
[73] Z. Gao, S. Zhang, K. Huang, Q. Wang, and C. Zhong, "Gradient distribution alignment certificates better adversarial domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8937–8946.
[74] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5018–5027.
[75] X. Peng et al., "VisDA: The visual domain adaptation challenge," 2017, *arXiv:1710.06924*.
[76] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 213–226.
[77] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb
[78] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 702–703.
[79] F. Zhu, Z. Cheng, X.-Y. Zhang, and C.-l. Liu, "Class-incremental learning via dual augmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 14306–14318.

Yukun Zuo (Graduate Student Member, IEEE) received the B.S. degree in information security in 2018 from the University of Science and Technology of China, Hefei, China, where he is currently working toward the Ph.D. degree. His research interests include domain adaptation and continual learning.



Hantao Yao (Member, IEEE) received the B.S. degree from XiDian University, Xi'an, China, in 2012, and the Ph.D. degree from the Institute of Computing Technology, University of Chinese Academy of Sciences, Beijing, China, in 2018. After graduation, he was a Postdoctoral from 2018 to 2020 with the National Laboratory of Pattern Recognition, Beijing, Institute of Automation, Chinese Academy of Sciences, Beijing. He is currently an Assistant Professor with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences. He was the recipient of the National Postdoctoral Programme for Innovative Talents. His research interests include zero-shot learning, person tracking and detection, and person re-identification.

Liansheng Zhuang (Member, IEEE) received the bachelor's and Ph.D. degrees from the University of Science and Technology of China (USTC), Hefei, China, in 2001 and 2006, respectively. In 2011, he was nominated to join the STARTRACKER Project of Microsoft Research of Asia, and he was a Vendor Researcher with the Visual Computing Group, Microsoft Research, Beijing, China. From 2012 to 2013, he was a Visiting Research Scientist with the Department of EECS, University of California at Berkeley, Berkeley, CA, USA. He is currently an Associate Professor with the School of Information Science and Technology, USTC. His main research interests include computer vision, and machine learning. He is a Member of ACM, and CCF.



Changsheng Xu (Fellow, IEEE) is currently a Professor with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Beijing, China, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He has hold 50 granted/pending patents and authored or coauthored more than 400 refereed research papers in these areas. His research interests include multimedia content analysis, pattern recognition, and computer vision. He was the Editor-in-chief, Associate Editor, Guest Editor, General Chair, Program Chair, Area/Track Chair, and TPC Member for more than 20 IEEE and ACM prestigious multimedia journals, conferences and workshops, including IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Multimedia Computing*, Communications and Applications and ACM Multimedia conference. He is IAPR Fellow and ACM Distinguished Scientist.