Adaptive Graph Embedding With Consistency and Specificity for Domain Adaptation

Shaohua Teng, *Member, IEEE*, Zefeng Zheng, Naiqi Wu, *Fellow, IEEE*, Luyao Teng, and Wei Zhang

***Abstract*—Domain adaptation (DA) aims to find a subspace, where the discrepancies between the source and target domains are reduced. Based on this subspace, the classifier trained by the labeled source samples can classify unlabeled target samples well. Existing approaches leverage Graph Embedding Learning to explore such a subspace. Unfortunately, due to 1) the interaction of the consistency and specificity between samples, and 2) the joint impact of the degenerated features and incorrect labels in the samples, the existing approaches might assign unsuitable similarity, which restricts their performance. In this paper, we propose an approach called adaptive graph embedding with consistency and specificity (AGE-CS) to cope with these issues. AGE-CS consists of two methods, i.e., graph embedding with consistency and specificity (GECS), and adaptive graph embedding (AGE). GECS jointly learns the similarity of samples under the geometric distance and semantic similarity metrics, while AGE adaptively adjusts the relative importance between the geometric distance and semantic similarity during the iterations. By AGE-CS, the neighborhood samples with the same label are rewarded, while the neighborhood samples with different labels are punished. As a result, compact structures are preserved, and advanced performance is achieved. Extensive experiments on five benchmark datasets demonstrate that the proposed method performs better than other Graph Embedding methods.**

***Index Terms*—Adaptive adjustment, consistency and specificity, domain adaptation, graph embedding, geometrical and semantic metrics.**

I. Introduction

A large amount of data from different domains is required to train a robust classification model. However, in some emerging target domains, only a small amount of labeled data is available, which is insufficient to learn critical classification knowledge. Moreover, it is time-consuming and costly to manually collect labeled data. In the light of these problems, domain adaptation (DA) is proposed to utilize labeled samples from a well-known domain (the source domain) to tag unlabeled samples from the emerging domain (the target domain) [1]. Up to now, DA has been widely applied to various fields, e.g., infection detection [2], [3], disease detection [4], anomaly detection [5]–[7], emotion recognition [8], and visual localization [9].

The primary nature of DA is to learn a projected subspace, where the discrepancies between the source and target domains are reduced [1], [10]. Based on a learned subspace, the classifier can properly classify the unlabeled target samples by utilizing the source knowledge.

Recently, some researchers adopt local structure preservation to align the distributions [11]–[13]. These methods construct a similarity matrix by measuring the geometric distance of samples, so as to preserve a local structure of the domains. However, here are still two issues to be addressed.

1) The Existing Methods Neglect the Interactions of the Consistency and Specificity Between Samples: The consistency denotes the common properties between samples, while specificity denotes specific properties of different samples. For example, the same category and common features of two samples might contribute to their consistency, while different categories and specific features of two samples might contribute to their specificity. In this case, there exist four possible relationships between two samples: a) a number of common features with the same category; b) a number of common features with different categories; c) a number of specific features with the same category; and d) a number of specific features with different categories.

Since most existing works measure similarity by geometric distance, *they might connect the samples a) and b) with larger weights, and the samples c) and d) with smaller weight*. As a result, the samples b) and c) are weighted inappropriately, and performance is limited. As revealed by [14], the consistency degree of a system reflects whether the projection is reliable or not. The result of having low consistency makes the knowledge of a model more unstable, which is not what we want. In order to achieve high consistency, an improved strategy should be used to measure the consistency and specificity between samples appropriately.

2) The Existing Methods Overlook Noise Samples That Con-

Manuscript received March 20, 2022; revised May 12, 2022 and July 21, 2022; accepted August 20, 2022. This work was supported in part by the Key-Area Research and Development Program of Guangdong Province (2020B010166006), the National Natural Science Foundation of China (61972102), the Guangzhou Science and Technology Plan Project (023A04J1729), and the Science and Technology development fund (FDCT), Macau SAR (015/2020/AMJ). Recommended by Associate Editor Xin Luo. *(Corresponding author: Luyao Teng.)*

Citation: S. H. Teng, Z. F. Zheng, N. Q. Wu, L. Y. Teng, and W. Zhang, "Adaptive graph embedding with consistency and specificity for domain adaptation," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 11, pp. 2094–2107, Nov. 2023.

S. H. Teng, Z. F. Zheng, and W. Zhang are with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China (e-mail: shteng@gdut.edu.cn; 2112005001@mail2.gdut.edu.cn; weizhang@gdut.edu.cn).

N. Q. Wu is with the Institute of Systems Engineering and Collaborative Laboratory for Intelligent Science and Systems, Macau University of Science and Technology, Macao 999078, China (e-mail: nqwu@must.edu.mo).

L. Y. Teng is with the School of Information Engineering, Guangzhou Panyu Polytechic, Guangzhou 511483, China, and also with the Faculty of Information Technology, Monash University, 20 Exhibition Walk Clayton, VIC 3800, Australia (e-mail: tengly@gzpyp.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2023.123318

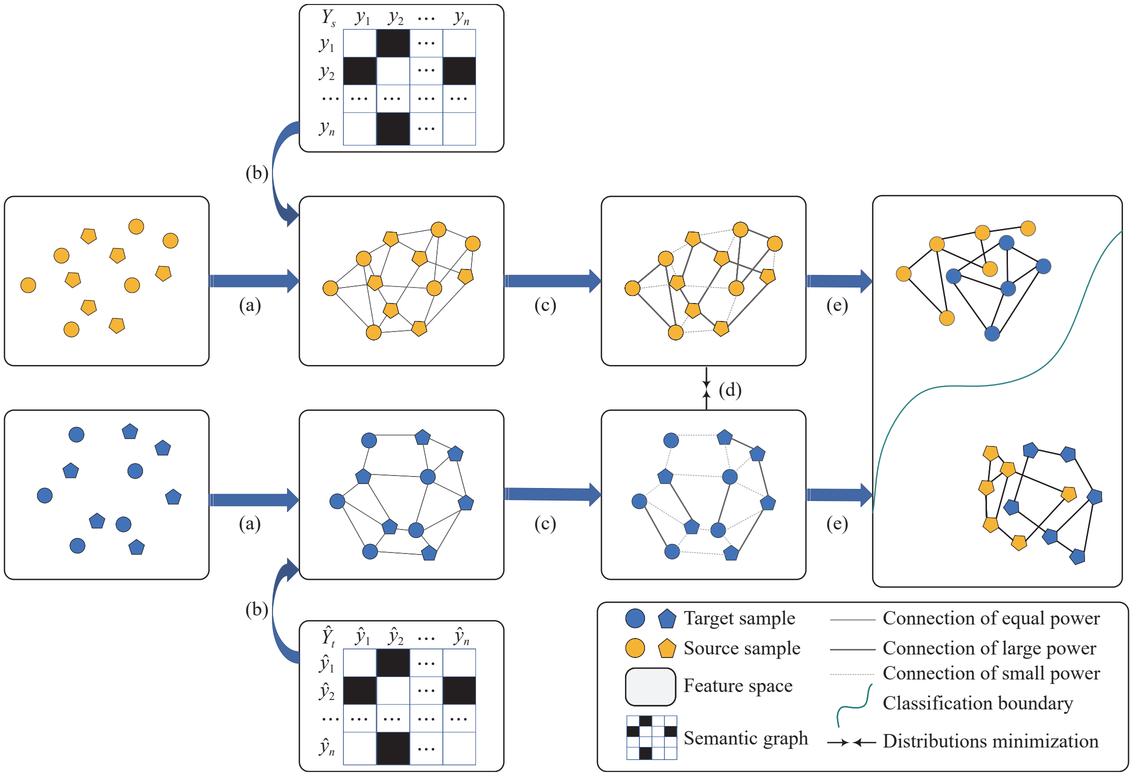

Y_s y_1 y_2 ⋯ y_n
y_1
y_2
y_n
(b)
(a)
(c)
(e)
(d)
(a)
(c)
(e)
(b)
$\hat{Y}_t$ $\hat{y}_1$ $\hat{y}_2$ ⋯ $\hat{y}_n$
$\hat{y}_1$
$\hat{y}_2$
$\hat{y}_n$
Target sample
Source sample
Feature space
Semantic graph
Connection of equal power
Connection of large power
Connection of small power
Classification boundary
Distributions minimization

Fig. 1. The flow chart of AGE-CS. (a) GECS measures the similarity of samples by geometric distance; (b) GECS measures the similarity of samples by semantic similarity metric; (c) AGE adaptively adjusts the relative importance of the geometric distance and semantic similarity metric; (d) MMD minimizes the distribution; and (e) By using AGE-CS, the compact structural information of domains is preserved, while discrepancies between domains are reduced. As a result, the discriminative classification boundary is obtained and advanced performance is guaranteed.

tain Degenerated Features or Incorrect Labels: In reality, there exist some samples with degenerated features [15] or incorrect labels [16]. These samples might mislead the similarity learning such that the local structure of domains cannot be well-preserved.

In light of the above two issues, one promising approach is to measure the similarity by the geometric distance and semantic information appropriately. However, it faces two challenges:

1) How can we unify the geometric distance and semantic similarity metrics to get a unified similarity?

2) How can we measure the relative importance between the geometric distance and semantic similarity metric, and adaptively adjust them?

In this paper, we address the above two issues, and introduce a novel method called adaptive graph embedding with consistency and specificity (AGE-CS). AGE-CS is composed of two parts: a) Graph embedding with consistency and specificity (GECS); and b) Adaptive graph embedding (AGE).

GECS adopts both the geometric distance and semantic similarity metrics to learn a similarity. In doing so, the neighborhood samples with same labels are rewarded, while the neighborhood samples with different labels are penalized. As a result, the consistency and specificity between samples are jointly measured, and promising performance is guaranteed.

AGE explores the potential relationship between geometry and semantics, and adaptively adjusts their weight based on the theoretical guarantee. By adopting AGE, the relative importance of the geometric distance and the semantic similarity metric is demonstrated. Hence, the structural information of two domains is preserved and advanced performance is achieved.

The contributions of this paper are as follows and the flowchart of AGE-CS is shown in Fig. 1.

1) AGE-CS is proposed, which consists of GECS and AGE. By AGE-CS, the compact structural information of domains is preserved, while the discrepancies between domains are reduced. As a result, advanced performance is achieved.

2) GECS jointly determines the similarity of samples under the geometric distance and semantic similarity metrics. Consequently, the neighborhood samples with same labels are rewarded, while the neighborhood samples with different labels are penalized.

3) AGE adaptively adjusts the relative importance between geometry and semantics, which results in compact structure preservation.

4) Extensive experiments and comparisons on five popular datasets are performed to demonstrate the effectiveness of the proposed method.

II. Related Work

In this section, we present a brief review of Graph Embedding methods, which can be divided into two categories, i.e., geometry-based graph embedding (GGE) methods [17]–[21]

and semantics-guided graph embedding (SGE) methods [12], [13], [22]–[25]. Interested readers can refer to the surveys [10] and [1] to gain a comprehensive perspective on DA methods.

As one of the categories, **approaches with geometry-based graph embedding (GGE)** assign the neighborhood relationship by feature matching and measure the similarity of samples by geometric distance. Liu *et al.* jointly adopt local and global GGE methods to explore the discriminative manifold structure of multi-source domains [17]. They hold the view that the distance between samples in the same domains is smaller than that in different domains. Thus, intra-class-and-inter-class-based GGE methods are proposed, and good performance is achieved on multi-source transfer tasks. Wang *et al.* propose manifold embedded distribution alignment (MEDA) that utilizes GGE to preserve the geometric structure of the learned manifold [18]. In MEDA, samples are projected into the manifold subspace, and their geometrical structures are explored simultaneously. As a result, MEDA avoids degenerated feature transformation and achieves promising performance. In addition, Vascon *et al.* apply an affinity matrix to convey the similarity of domains [19]. Since they propagate the similarity between the labels directly, the target labels are obtained effectively. Moreover, Xiao *et al.* leverage both low-rank representation and GGE to preserve the structural relationships of samples [20]. They jointly explore the discriminative features of samples and label information. With the help of τ-technology, a linear regression classifier is achieved and the geometric structure is mined. Differently, Sun *et al.* jointly utilize the maximum mean discrepancy (MMD), manifold learning, and scatter preservation to learn discriminative and domain-invariant features [21]. During training, semantics and features are incorporated into a latent example-class matrix, and the geometrical information is explored on the latent space. With experiments, they verify the effectiveness of GGE.

As the other category, **semantics-guided graph embedding (SGE) methods** assign the neighborhood relationship by semantic mapping and measure the similarity of samples by a geometric metric. If all samples are connected, SGE is equivalent to the scatter component analysis (SCA) [26]. Li *et al.* propose domain invariant and class discriminative (DICD) that jointly adopts within-class and between-class scatters to learn domain-invariant features [22]. Since both intra-class and inter-class SGE methods are employed, DICD digs out the discriminative information sufficiently and achieves compact clusters. Li *et al.* embed SGE into a coupled projection learning framework [23]. The distributions, scatters, and semantics are jointly leveraged, and a more feasible solution is gained by solving two coupled projection matrices. Gholenji and Tahmoresnezhad adopt both distribution alignment and discriminative manifold learning methods to exploit statistical, local, and global structures [13]. Different from DICD, Gholenji *et al.* introduce repulsive terms to align cross-domain distributions, which leads to consistent representation. However, since additional constraints are involved, this method requires more training time. Zhao *et al.* use density peak landmark selection (DPLS) and manifold learning to mine the potential structural information of domains [24]. In this way, samples are well-measured according to global density. By DPLS, the reliable samples are selected and the geometric structures are further explored by these high-quality samples. In experiments, they verify its significant improvements. Meng *et al.* jointly preserve the marginal and local structures to obtain discriminant information and propose margin and locality structure preservation [12]. Different from SPDA, it focuses on exploring consistent and inconsistent information, which exhibits promising performance in the few-shot setting. Li *et al.* propose Label Correction to align the distribution shift caused by the target pseudo labels [25]. Based on the SGE method, they divide the optimization process into two stages. At the first stage, they align the distributions by minimizing marginal and conditional distributions. Then, they correct the target pseudo labels so as to further align the distributions. Since distributions are well-measured on these two stages, their method achieves significant performance. However, it takes more time to align the distributions.

Although the above-mentioned methods achieve promising improvements, they not only neglect the interactions of the consistency and specificity between samples, but also do not consider noise samples. As a result, further research is necessary.

Different from the previous works, in light of the unsolved problems, in this paper, we measure the similarity under both geometric distance and semantic similarity. The differences between the proposal and previous works are two folds.

1) We propose GECS to measure the similarity of samples from both geometric distance and semantic similarity perspectives, while the afore-mentioned studies cope with one of them only. By using GECS, the consistent and specific properties of domains are further explored and performance is improved.

2) We propose AGE to adaptively measure the relative importance between geometry and semantics. A mathematical analysis of the optimal parameter is given (refer to Theorem 1) and the transfer performance is guaranteed. To our best knowledge, there is no relevant study that reveals the relative importance between geometry and semantics mathematically.

III. Proposed Method

This section introduces AGE-CS in detail. First, we give the notations used in this paper and the problem setting. Then, the conventional methods and their drawbacks are reviewed. Next, the proposed GECS and AGE are discussed. At last, the overall objective function and its optimization procedure are given.

A. Notations and Problem Setting

In this subsection, the notations used in this paper are shown in Table I, and the problem setting of DA is as follows.

Problem Setting: Let $X_s = \{x_{s,i}\}_{i=1}^{n_s}$ be the set of source samples, $X_t = \{x_{t,i}\}_{i=1}^{n_t}$ be the set of target samples, and $Y_s =$

TABLE I
THE NOTATIONS

Notation	Dimension	Description
$\mathcal{P}_{X_s,Y_s}$	–	The joint distribution of X_s
$\mathcal{P}_{X_t,Y_t}$	–	The joint distribution of X_t
d	$\mathbb{R}$	The dimension after projection
k	$\mathbb{R}$	The neighborhood number
W	$\mathbb{R}$	The projection matrix
m	$\mathbb{R}$	The dimension of a domain
n	$\mathbb{R}$	The number of source and target samples ($n = n_s + n_t$)
n_s	$\mathbb{R}$	The number of source samples
n_t	$\mathbb{R}$	The number of target samples
$(\beta_s)_i$	$\mathbb{R}$	The importance weight of the semantic information of $x_{s,i}$
$(\beta_t)_i$	$\mathbb{R}$	The importance weight of the semantic information of $x_{t,i}$
γ_s	$\mathbb{R}$	The regularization parameter of S^s
γ_t	$\mathbb{R}$	The regularization parameter of S^t
$\alpha, \lambda, \delta, \tau$	$\mathbb{R}$	The hyper-parameters
$\mathbf{1}$	$\mathbb{R}^n$	The row vector with all elements being one
Y_s	$\mathbb{R}^{n_s}$	The source label matrix
$\hat{Y}_t$	$\mathbb{R}^{n_t}$	The target pseudo label matrix
X_s	$\mathbb{R}^{m\times n_s}$	The set of source samples
X_t	$\mathbb{R}^{m\times n_t}$	The set of target samples
S_s	$\mathbb{R}^{n_s\times n_s}$	The similarity matrix of the source samples
S_t	$\mathbb{R}^{n_t\times n_t}$	The similarity matrix of the target samples
G_s	$\mathbb{R}^{n_s\times n_s}$	The semantic graph of X_s
G_t	$\mathbb{R}^{n_t\times n_t}$	The semantic graph of X_t
$\hat{G}_s$	$\mathbb{R}^{n_s\times n_s}$	The semantic graph of X_s organized by distance from small to large
$\hat{G}_t$	$\mathbb{R}^{n_t\times n_t}$	The semantic graph of X_t organized by distance from small to large
H	$\mathbb{R}^{n\times n}$	The centering matrix

$\{y_{s,i}\}_{i=1}^{n_s} \in \mathbb{R}^{n_s}$ be the set of source labels, where n_s and n_t are the number of source and target samples, respectively. DA assumes that the source domain $D_s = \{(x_{s,i}, y_{s,i})\}_{i=1}^{n_s}$ comes from the source joint distributions $\mathcal{P}_{X_s,Y_s}$, while the target domain $D_t = \{(x_{t,i})\}_{i=1}^{n_t}$ is from the target joint distribution $\mathcal{P}_{X_t,Y_t}$, where $\mathcal{P}_{X_s,Y_s} \neq \mathcal{P}_{X_t,Y_t}$ and $Y_t \in \mathbb{R}^{n_t}$ is inaccessible during training. Then, DA aims to learn a projection matrix $W \in \mathbb{R}^{m\times d}$ such that the classifier f trained on the source domain can classify the target domain correctly. That is,

$$\min_W \mathbb{L}(f(W^T X_t), Y_t)) \tag{1}$$

where $\mathbb{L}(A, B)$ is a loss function that measures the loss between A and B, $f(X)$ is the classifier trained on feature space X, and Y_t is the set of ground-truth labels of X_t, which is inaccessible during training.

B. Distribution Alignment and Graph Embedding

In this subsection, we review conventional methods and point out two existing problems.

In order to reduce the discrepancies between two domains, the usual practice is to reduce the marginal and conditional distributions between the two domains [27]. That is,

$$\begin{aligned}&\min_W f_{\mathrm{DA}}(W, X_s, X_t)\\&=\mu \left\| \frac{1}{n_s}\sum_{i=1}^{n_s} W^T x_{s,i} - \frac{1}{n_t}\sum_{j=1}^{n_t} W^T x_{t,j} \right\|_F^2\\&\quad + (1-\mu)\sum_{c=1}^{C} \left\| \frac{1}{n_s^c}\sum_{i=1}^{n_s^c} W^T x_{s,i}^c - \frac{1}{n_t^c}\sum_{j=1}^{n_t^c} W^T x_{t,j}^c \right\|_F^2\\&\text{s.t. } W^T X H X^T W = I\end{aligned} \tag{2}$$

where

1) $W \in \mathbb{R}^{m\times d}$ is the projection matrix;

2) $X_s^c = \{x_{s,i}\}_{i=1}^{n_s^c}$ and $X_t^c = \{x_{t,i}\}_{i=1}^{n_t^c}$ denote the set of samples belonging to c-th class of the source and target domains, respectively;

3) n_s^c represents the number of elements in X_s^c, and n_t^c represents the number of elements in X_t^c;

4) $H = I_{n\times n} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the centering matrix that maximizes the scatter of samples, $I_{n\times n} \in \mathbb{R}^{n\times n}$ is an identity matrix, and $\mathbf{1} \in \mathbb{R}^n$ is a vector with all elements in it being 1;

5) μ is a hyper-parameter.

However, by adopting (2), the discrepancies between the two domains might still be large, which degrades performance. Hence, the local connectivity of each domain is explored by using Graph Embedding as follows.

Definition 1 (Graph embedding (GE)): Given sample domain $X = \{x_i\}_{i=1}^{n} \in \mathbb{R}^{m\times n}$, GE is defined as the actions $f_{\mathrm{GE}}(W, S, \gamma, X)$ that preserve the structure of domain X by learning a projection matrix $W \in \mathbb{R}^{m\times d}$ and a similarity matrix $S^{n\times n}$. That is,

$$\begin{aligned}&\min_{W,S,\gamma} f_{\mathrm{GE}}(W, S, \gamma, X)\\&= \sum_{i=1}^{n}\sum_{j=1}^{n} \left\|W^T x_i - W^T x_j\right\|_2^2 s_{i,j} + \gamma s_{i,j}^2\\&\text{s.t. } \sum_{j=1}^{n} s_{i,j} = 1, 0 \leq s_{i,j} \leq 1\end{aligned} \tag{3}$$

where γ is a parameter to be learned.

Remark 1: It is worth noting that Definition 1 is slightly different from the conventional one [28]. However, Definition 1 unifies the definitions used in DA [12], [19], [21], [25], [28]–[30] by the following three updated strategies:

a) S is given by graph mapping and is fixed during the optimization process [12], [19], [21], [28], [29], [31]. In this case, two factors need to be ensured: reliable semantic mapping and accurate feature measurement;

b) S is given by graph mapping and is updated during the optimization process [30], [32]. In this case, three factors need to be ensured: reliable semantic mapping, accurate feature measurement, and reliable information extraction in the iterative process;

c) S is learned according to its constraints during the optimization process [25]. In this case, four factors should be guaranteed: reliable semantic mapping, accurate feature measurement, reliable information extraction in the iterative process, and stable convergence.

Remark 2: Since strategy c) is more challenging than strategies a) and b), in this paper, we discuss GE with strategy c). The methods that apply the above strategies are compared in the experiments.

Based on (3), we simultaneously reduce the discrepancies between the domains and explore the local structures of the source and target domains by (4). That is,

$$\begin{aligned}\min_{W,S_s,S_t,\gamma_s,\gamma_t} & f_{\mathrm{DA}}(W,X_s,X_t)\\ &+f_{\mathrm{GE}}(W,S_s,\gamma_s,X_s)\\ &+f_{\mathrm{GE}}(W,S_t,\gamma_t,X_t)\end{aligned}\tag{4}$$

where $S_s=\{(s_s)_{i,j},1\le i,j\le n_s\}\in\mathbb{R}^{n_s\times n_s}$ and $S_t=\{(s_t)_{i,j},1\le i,j\le n_t\}\in\mathbb{R}^{n_t\times n_t}$ are the similarity matrices of the source and target domains, respectively.

With the above actions, the geometric structures of the two domains are explored and the discrepancies between the two domains are reduced. As a result, joint knowledge of the features is learned and good performance is obtained.

Unfortunately, there are still two factors that might hinder the performance:

1) The Interactions of the Consistency and Specificity Between Samples: In (4), the similarity matrices S_s and S_t are formed by measuring the geometric distances on the subspace W^TX, respectively. However, these actions cannot be performed in these two mentioned situations, i.e., a) two samples share a number of common features in different categories; and b) two samples share a number of specific features in the same category.

2) The Samples With Degenerated Features in the Two Domains: There might exist some samples with degenerated features in the two domains [15], which distorts the subspace learning. If the learned feature space W^TX is distorted, the similarity of samples might be wrongly measured and performance is hindered.

For these problems, we propose GECS in the next subsection.

C. Graph Embedding With Consistency and Specificity

To overcome the drawbacks of (4), we introduce GECS as presented by Definition 2.

Definition 2 (Graph embedding with consistency and specificity (GECS)): Let $X=\{x_i\}_{i=1}^n\in\mathbb{R}^{m\times n}$ be a sample set and $Y\in\mathbb{R}^n$ be the label matrix of X. Given semantic graph $G\in\mathbb{R}^{n\times n}$ with each element $g_{i,j}=\begin{cases}1, & y_i=y_j\\ 0, & y_i\ne y_j\end{cases}$, GECS is defined as the action $f_{\mathrm{GECS}}(W,S,\gamma,\Upsilon,G,X)$ that embeds the semantic graph G into the learning process of GE $f_{\mathrm{GE}}(W,S,\gamma,X)$. That is,

$$\begin{aligned}\min_{W,S,\gamma,\Upsilon} & f_{\mathrm{GECS}}(W,S,\gamma,\Upsilon,G,X)\\ &=\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|W^Tx_i-W^Tx_j\right\|_2^2 s_{i,j}\\ &\quad+\|\Upsilon(S-G)\|_F^2+\gamma s_{i,j}^2\\ &\text{s.t.}\sum_{j=1}^{n}s_{i,j}=1,0\le s_{i,j}\le 1\end{aligned}\tag{5}$$

where $\Upsilon=\begin{bmatrix}\sqrt{\beta_1} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{\beta_n}\end{bmatrix}$ is a matrix of hyper-parameters for balancing the relative importance of the geometric and semantic distances of each sample x_i.

By (5), the objective function given by (4) is modified to

$$\begin{aligned}\min_{W,S_s,S_t,\gamma_s,\gamma_t,\Upsilon_s,\Upsilon_t} & f_{\mathrm{DA}}(W,X_s,X_t)\\ &+f_{\mathrm{GECS}}(W,S_s,\gamma_s,\Upsilon_s,G_s,X_s)\\ &+f_{\mathrm{GECS}}(W,S_t,\gamma_t,\Upsilon_t,G_t,X_t)\end{aligned}\tag{6}$$

where

1) $G_s\in\mathbb{R}^{n_s\times n_s}$ and $G_t\in\mathbb{R}^{n_t\times n_t}$ are the semantic graph of the source and target domains with each element

$$(g_s)_{i,j}=\begin{cases}1, & y_{s,i}=y_{s,j}\\ 0, & y_{s,i}\ne y_{s,j}\end{cases}\tag{7}$$

and

$$(g_t)_{i,j}=\begin{cases}1, & \hat{y}_{t,i}=\hat{y}_{t,j}\\ 0, & \hat{y}_{t,i}\ne\hat{y}_{t,j}.\end{cases}\tag{8}$$

2) $\Upsilon_s\in\mathbb{R}^{n_s\times n_s}$ and $\Upsilon_t\in\mathbb{R}^{n_t\times n_t}$ are matrices of hyper-parameters of the source and target domains, respectively, i.e.,

$$\Upsilon_s=\begin{bmatrix}\sqrt{(\beta_s)_{11}} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{(\beta_s)_{n_s}}\end{bmatrix}\tag{9}$$

and

$$\Upsilon_t=\begin{bmatrix}\sqrt{(\beta_t)_1} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{(\beta_t)_{n_t}}\end{bmatrix}.\tag{10}$$

3) $\hat{y}_{t,i}\in\hat{Y}_t$ is the pseudo label of $x_{t,i}$.

With these actions, two advantages are obtained:

1) GECS Compensates for Inappropriate Assignment Caused by the Geometric Distance: Due to the interaction of the consistency and specificity between samples, the neighborhood samples with different labels might be connected with a large weight, which breaks the concept of similarity learning. By introducing GECS, the neighborhood samples with the same label are rewarded, while the neighborhood samples with different labels are punished. By doing so, the similarity is jointly measured.

2) GECS Reduces the Impact of Noise Samples: Since (4) measures the similarity by the geometric distance, it might be affected by samples with degenerated features. In (6), GECS embeds the semantic information into the similarity learning and jointly measures the similarity, which corrects the inappropriate measurement caused by noise samples.

As a result, the similarities are remeasured under both geometric distance and semantic similarity metrics, and a compact structure is guaranteed. Apart from the promising performance of GECS, another problem catches our attention: *Determine how to measure the relative importance between the geometric distance and semantic similarity of each sample.*

Due to the involvement of the semantic graphs G_s and G_t, the strategy proposed in [33], [34] does not work. In this case, further work should be done. In the next subsection, we propose AGE to adaptively adjust the hyper-parameters of GECS.

D. Adaptive Graph Embedding

To adaptively adjust the hyper-parameters in GECS, we introduce AGE as given by Definition 3. Then, Theorem 1 is proposed to guarantee the optimal solutions of AGE.

Definition 3 (Adaptive graph embedding (AGE)): Let $X = \{x_i\}_{i=1}^n \in \mathbb{R}^{m\times n}$ be a sample set, $Y \in \mathbb{R}^n$ be the label matrix of X, and $G \in \mathbb{R}^{n\times n}$ be a semantic graph with each element $g_{i,j} = \begin{cases} 1, & y_i = y_j \\ 0, & y_i \neq y_j \end{cases}$. AGE is defined as the processes to update the balance parameter β_i and regularization parameter γ in (5) by the updating strategies $\mathcal{G}$ and $\mathcal{F}$ in $\mathcal{T}$-th iteration, respectively, i.e.,

$$\beta_i^{\mathcal{T}} = \mathcal{G}(X, G^{\mathcal{T}}), \forall i \in [1, n] \tag{11}$$

$$\gamma^{\mathcal{T}} = \mathcal{F}(X, G^{\mathcal{T}}, \beta_i^{\mathcal{T}}), \forall i \in [1, n]. \tag{12}$$

The performance of the updating strategies $\mathcal{G}$ and $\mathcal{F}$ is theoretically guaranteed based on the following theorem.

Theorem 1: Let $W \in \mathbb{R}^{m\times d}$ be a projection matrix, $X \in \mathbb{R}^{m\times n}$ be a sample set, and $G \in \mathbb{R}^{n\times n}$ be a supervised graph with each element $g_{i,j} = \begin{cases} 1, & y_i = y_j \\ 0, & y_i \neq y_j \end{cases}$, where $y_i \in Y$ is the semantics of sample set X. To solve the optimization problem given by (11) and (12), parameter β_i can be updated by

$$\beta_i = \begin{cases} -\dfrac{\hat{d}_{i,k+1} - \hat{d}_{i,k}}{k(\hat{g}_{i,k+1} - \hat{g}_{i,k})}, & \hat{g}_{i,k+1} \neq \hat{g}_{i,k} \\ \tau, & \hat{g}_{i,k+1} = \hat{g}_{i,k} \end{cases} \tag{13}$$

while parameter γ can be updated by

$$\gamma = \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{2}(k\hat{d}_{i,k} - \sum_{j=1}^{k}\hat{d}_{i,j}) - \beta_i(k\hat{g}_{i,k} - \sum_{j=1}^{k}\hat{g}_{i,j} + 1)) \tag{14}$$

where

1) k is the neighborhood number;

2) $\hat{x}_j$ is the j-th nearest sample from x_i;

3) $d_{ij} = \|W^T x_i - W^T x_j\|_2^2$ is the distance metric of samples x_i and x_j on the subspace $W^T X$;

4) $\hat{d}_{i,j}$ denotes the distance between x_i and $\hat{x}_j$, and $\hat{d}_{i,:} = \{\hat{d}_{i,1}, \hat{d}_{i,2}, \ldots, \hat{d}_{i,n}\}$ is the sorted vector of x_i that arranges $d_{i,:} = \{d_{i,1}, d_{i,2}, \ldots, d_{i,n}\}$ in a non-decreasing order;

5) $\hat{g}_{i,j}$ denotes the semantic relationship of x_i and $\hat{x}_j$, i.e.,

$$\hat{g}_{i,j} = \begin{cases} 1, & \text{if } x_i \text{ and } \hat{x}_j \text{ share the same label} \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

6) $\tau \in \mathbb{R}$ is an arbitrary value that is used to emphasize the semantic information;

Proof: The proof can be found in Appendix. ■

From Theorem 1, (11) and (12) are learned by (13) and (14), respectively. Hence, the hyper-parameters $(\beta_s)_i$ and $(\beta_t)_j$ $(1 \leq i \leq n_s, 1 \leq j \leq n_t)$ in (6) can be updated by (13), while the hyper-parameters γ_s and γ_t can be updated by (14), respectively. By doing so, the relative importance between the geometric distance and semantic similarity is well-measured during the iteration and the impact of the noise samples is further eased. Therefore, advanced performance is achieved.

E. Overall Objective Function

In this subsection, we give the objective function of AGE-CS. By bringing (2) and (5) to (6), the objective function of the proposed method can be written in a matrix form as

$$\begin{aligned} \min_{W,S,\Upsilon,\gamma_s,\gamma_t} \ & f_{\mathrm{DA}}(W, X_s, X_t) \\ & + f_{\mathrm{GECS}}(W, S_s, \gamma_s, \Upsilon_s, G_s, X_s) \\ & + f_{\mathrm{GECS}}(W, S_t, \gamma_t, \Upsilon_t, G_t, X_t) \\ = \ & tr(W^T X(M + \alpha L)X^T W) + + \|\Upsilon(S - G)\|_F^2 \\ & + \gamma_s (s_s)_{i,j}^2 + \gamma_t (s_t)_{i,j}^2 \\ \text{s.t.} \ & \sum_{j=1}^{n_s}(s_s)_{i,j} = 1, 0 \leq (s_s)_{i,j} \leq 1, \forall i, j \in [1, n_s] \\ & \sum_{j=1}^{n_t}(s_t)_{i,j} = 1, 0 \leq (s_t)_{i,j} \leq 1, \forall i, j \in [1, n_t] \\ & W^T X H X^T W = I \end{aligned} \tag{16}$$

where

1) $X = [X_s, X_t]$ denotes the sample set of the two domains;

2) M_0 is a matrix that measures the marginal distribution between X_s and X_t by (17), while $\sum_{c=1}^{C} M_c$ computes their conditional distribution according to (18). By combining the above matrices together, $M = \mu M_0 + (1-\mu)\sum_{i=1}^{C} M_c$ measures the marginal and conditional distributions jointly;

$$(M_0)_{i,j} = \begin{cases} \dfrac{1}{n_s^2}, & x_i, x_j \in X_s \\ \dfrac{1}{n_t^2}, & x_i, x_j \in X_t \\ -\dfrac{1}{n_s n_t}, & \text{otherwise} \end{cases} \tag{17}$$

$$(M_c)_{i,j} = \begin{cases} \dfrac{1}{(n_s^c)^2}, & x_i, x_j \in X_s^c \\ \dfrac{1}{(n_t^c)^2}, & x_i, x_j \in X_t^c \\ -\dfrac{1}{n_s^c n_t^c}, & \begin{cases} x_i \in X_s^c \wedge x_j \in X_t^c \\ x_j \in X_s^c \wedge x_i \in X_t^c \end{cases} \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

3) $L = D - S \in \mathbb{R}^{n\times n}$ is the Laplacian matrix, $D \in \mathbb{R}^{n\times n}$ is the degree matrix whose diagonal element is calculated by $D_{i,i} = \sum_{j=1}^{n} s_{i,j}$, and $S = \begin{bmatrix} S_s & 0 \\ 0 & S_t \end{bmatrix}$ is the learned similarity matrix of the two domains; and

4) $\Upsilon = \begin{bmatrix} \Upsilon_s & 0 \\ 0 & \Upsilon_t \end{bmatrix}$ is a group of parameters of X, which gives the weight of the semantics.

F. Optimization Process

According to (16), there are six variables, i.e., Υ_s, Υ_t, S, W, γ_s, and γ_t that need to be optimized.

Before presenting the optimization process, we give the following symbols.

1) $\hat{x}_{s,j}$ is the j-th nearest source sample from $x_{s,i}$;

2) $\hat{x}_{t,j}$ is the j-th nearest target sample from $x_{t,i}$;

3) $(\hat{d}_s)_{i,j} = \|x_{s,i} - x_{s,j}\|_2^2$ is the distance between $x_{s,i}$ and $\hat{x}_{s,j}$;

4) $(\hat{d}_t)_{i,j} = \|x_{t,i} - x_{t,j}\|_2^2$ is the distance between $x_{t,i}$ and $\hat{x}_{t,j}$;

5) $\hat{G}_s \in \mathbb{R}^{n_s\times n_s}$ is the sorted semantic graph of X_s with each element giving as

$$(\hat{g}_s)_{i,j} = \begin{cases} 1, & \text{if } x_{s,i} \text{ and } \hat{x}_{s,j} \text{ share the same label} \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

6) $\hat{G}_t \in \mathbb{R}^{n_t\times n_t}$ is the sorted semantic graph of X_t with each element giving as

$$(\hat{g}_t)_{i,j} = \begin{cases} 1, & \text{if } x_{t,i} \text{ and } \hat{x}_{t,j} \text{ share the same pseudo label} \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Then, we update each of the variables alternatively while keeping the others fixed.

1) Update Υ_s and Υ_t: According to (13), $(\beta_s)_i$ and $(\beta_t)_i$ are updated by

$$(\beta_s)_i = \begin{cases} \dfrac{(\hat{d}_s)_{i,k+1} - (\hat{d}_s)_{i,k}}{k((\hat{g}_s)_{i,k+1} - (\hat{g}_s)_{i,k})}, & (\hat{g}_s)_{i,k+1} \neq (\hat{g}_s)_{i,k} \\ \tau, & (\hat{g}_s)_{i,k+1} = (\hat{g}_s)_{i,k} \end{cases} \tag{21}$$

$$(\beta_t)_i = \begin{cases} \dfrac{(\hat{d}_t)_{i,k+1} - (\hat{d}_t)_{i,k}}{k((\hat{g}_t)_{i,k+1} - (\hat{g}_t)_{i,k})}, & (\hat{g}_t)_{i,k+1} \neq (\hat{g}_t)_{i,k} \\ \tau, & (\hat{g}_t)_{i,k+1} = (\hat{g}_t)_{i,k} \end{cases} \tag{22}$$

where τ is an arbitrary value that is used to emphasize semantic information.

2) Update γ_s and γ_t: According to (14), γ_s and γ_t are updated by

$$\begin{aligned} \gamma_s = &\frac{1}{n_s}\sum_{i=1}^{n_s}\Big(\frac{1}{2}\big(k(\hat{d}_s)_{i,k} - \sum_{j=1}^{k}(\hat{d}_s)_{i,j}\big) \\ &- (\beta_s)_i \times \big(k(\hat{g}_s)_{i,k} - \sum_{j=1}^{k}(\hat{g}_s)_{i,j} + 1\big)\Big) \end{aligned} \tag{23}$$

$$\begin{aligned} \gamma_t = &\frac{1}{n_t}\sum_{i=1}^{n_t}\Big(\frac{1}{2}\big(k(\hat{d}_t)_{i,k} - \sum_{j=1}^{k}(\hat{d}_t)_{i,j}\big) \\ &- (\beta_t)_i \times \big(k(\hat{g}_t)_{i,k} - \sum_{j=1}^{k}(\hat{g}_t)_{i,j} + 1\big)\Big). \end{aligned} \tag{24}$$

3) Update S: Updating S is the same as updating S_s and S_t. According to (32), S_s and S_t can be updated row by row with their k-nearest elements $\hat{x}_{s,j}$ and $\hat{x}_{t,j}$, respectively, i.e.,

$$(\hat{s}_s)_{i,j} = \frac{1}{\gamma_s + (\beta_s)_i}\Big(-\frac{1}{2}(\hat{d}_s)_{i,j} + (\beta_s)_i \times (\hat{g}_s)_{i,j}\Big), \forall j \in [1,k] \tag{25}$$

$$(\hat{s}_t)_{i,j} = \frac{1}{\gamma_t + (\beta_t)_i}\Big(-\frac{1}{2}(\hat{d}_t)_{i,j} + (\beta_t)_i \times (\hat{g}_t)_{i,j}\Big), \forall j \in [1,k]. \tag{26}$$

Once S_s and S_t are computed, we smooth S in the $\mathcal{T}$-th iteration as

$$S^{\mathcal{T}} = \delta \begin{bmatrix} (S_s)^{\mathcal{T}} & 0 \\ 0 & (S_t)^{\mathcal{T}} \end{bmatrix} + (1-\delta)S^{\mathcal{T}-1} \tag{27}$$

where δ is a hyper-parameter that smooths the learning of S.

4) Update W: Let the derivative of (16) with respect to W be zero, the solution can be derived as a generalized eigenvalue-decomposition problem as

$$(X(M + \alpha L)X^T + \lambda I_{m\times m})W = XHX^T W\Theta \tag{28}$$

where $I_{m\times m} \in \mathbb{R}^{m\times m}$ is an identity matrix that avoids the trivial solution and λ is a hyper-parameter.

For a better illustration, we summarize the algorithm procedure as shown in **Algorithm 1**.

Algorithm 1 Adaptive Graph Embedding With Consistency and Specificity (AGE-CS)

```
Input:    $X_s$ and $X_t$: the source and target samples;
          $Y_s$: the source label;
          $d$: the dimensionality of the projection subspace;
          $\alpha, \lambda, \delta, k, \tau$: the hyper-parameters;
          $T$: the number of iterations;
Output:   $W$: the projection matrix;
          $\hat{Y}_t$: the pseudo label matrix for target samples $X_t$
Initialize $\widetilde{T} = 0$; $S = 0$;
Initialize pseudo label $\hat{Y}_t$ by training $X_s$ and $X_t$;
while: $\widetilde{T} < T$ do
    1) $\widetilde{T} \leftarrow \widetilde{T} + 1$;
    2) Update $(\beta_s)_i$ and $(\beta_t)_i$ by (21) and (22), respectively;
    3) Update $\gamma_s$ and $\gamma_t$ by (23) and (24), respectively;
    4) Update $S$ by (25)–(27);
    5) Update $W$ by (28);
    6) Train classifier $f$ by $W^T X_s$ and $W^T X_t$;
    7) Update $\hat{Y}_t$ by the trained classifier $f$.
end while
```

G. Time Complexity

The complexity of the components during the optimization process is as follows:

a) The complexity of constructing matrix L_s and L_t is $O(n_s^2)$ and $O(n_t^2)$, respectively;

b) The complexity of constructing matrix M_0 is $O(n^2)$;

c) The complexity of constructing matrix $\sum_{c=1}^{C} M_c$ is $O(n^2)$;

d) The complexity of solving the generalized eigenvalue-decomposition problem with respect to (28) is $O(m^3)$.

TABLE II
OVERVIEW OF THE DATASETS

Dataset	Domain	Feature	Class	Sample
Office+Caltech10	A, D, C, W	SURF (800)	10	958/157/1123/295
Office31	A, D, W	ResNet50 (2048)	31	2817/498/795
Office-Home	Ar, Cl, Pr, Re	ResNet50 (2048)	65	2421/4379/4428/4357
ImageCLEF-DA	C, I, P	ResNet50 (2048)	12	600/600/600
COIL20	COIL1, COIL2	Raw (1024)	20	720/720

Assume T is the number of iterations, and the overall complexity of AGE-CS is

$$O(T(n_s^2+n_t^2+2n^2+m^3+m^2n))=O(T(n^2+m^2n+m^3)).$$

IV. Experiments

In this section, we first describe the five involved datasets and the experimental settings. Then, comparison experiments with other popular algorithms are given. Moreover, the parameter sensitivity, convergence analysis, and ablation experiments of AGE-CS are evaluated. For the sake of reproduction, the source codes for the experiments are released at https://github.com/zzf495/AGE-CS. Besides, we introduce a promising repository that implements some of the shallow domain adaptation methods at https://github.com/zzf495/Re-implementations-of-SDA.

A. Involved Datasets

In this paper, we adopt five widely used databases for experiments, including Office+Caltech10, Office31, Office-Home, ImgeCLEF-DA, and COIL20. The overview of the datasets are shown in Table II and the details are as follows.

Office+Caltech10 [35] is a commonly used dataset in shallow transfer learning. It includes four sub-domains, i.e., A (Amazon), W (Webcam), C (Caltech), and D (DSLR). The number of images for Amazon, Webcam, Caltech, and DSLR is 958, 295, 1123, and 157, respectively, with each domain containing 10 classes. In the experiments, we use the SURF features extracted by [35], where the images with 800-bin histograms are trained and encoded. Interested readers can refer to [35] for details of data processing. 12 cross-domain tasks, e.g., $\mathrm{A}\rightarrow\mathrm{D}$, $\mathrm{A}\rightarrow\mathrm{W}$, $\mathrm{A}\rightarrow\mathrm{C},\ldots,$ and $\mathrm{C}\rightarrow\mathrm{W}$ are conducted for comparisons.

Office31 [36] is composed of three sub-domains, i.e., Amazon (A), DSLR (D), and Webcam (W). The dataset is formed from 4110 images of objects in 31 common categories. Its sub-domains are composed of online e-commerce pictures, high-resolution pictures, and low-resolution pictures, respectively. In the experiments, features with 2048 dimensions are extracted by using ResNet50 and six tasks are formed, i.e., $\mathrm{A}\rightarrow\mathrm{D}, \mathrm{A}\rightarrow\mathrm{W},\ldots, \mathrm{W}\rightarrow\mathrm{D}$.

Office-Home [37] contains 65 kinds of different objects with 30 475 original samples and is composed of four sub-domains: Art (Ar), Clipart (Cl), Product (Pr), and Real-World (Re). The sizes of these sub-domains are 2427, 4365, 4439, and 4357, respectively. In the experiments, we use ResNet50 models to extract the features and conduct 12 cross-domain tasks, i.e., $\mathrm{Ar}\rightarrow\mathrm{Cl}$, $\mathrm{Ar}\rightarrow\mathrm{Pr}$ $,\ldots,\mathrm{Re}\rightarrow\mathrm{Pr}$.

ImageCLEF-DA includes three sub-domains, i.e., Caltech-256 (C), ImageNet ILSVRC2012(I), and Pascal VOC2012 (P). Each domain contains 600 images of 12 categories with 2,048 dimensions. Following [32], six cross-domain tasks, i.e., $\mathrm{C}\rightarrow\mathrm{I}$, $\mathrm{C}\rightarrow\mathrm{P},\ldots,\mathrm{P}\rightarrow\mathrm{I}$, are performed in the experiments.

COIL20 [38] consists of two domains, i.e., COIL1 and COIL2, with 1440 images in each domain. The dataset is formed by taking 75 images as the base and deriving new images every five degrees of rotation. COIL1 contains the images in $[0°,85°]\cup[180°,265°]$, while COIL2 contains the images in $[90°,175°]\cup[270°,365°]$. In the experiments, two cross-domain tasks are adopted, i.e., COIL1 → COIL2 and COIL2 → COIL1.

B. Comparison Method

For comparisons, seven state-of-the-art shallow transfer learning methods are introduced:

Domain Invariant and Class Discriminative Feature Learning (DICD, 2018) [22] which jointly minimizes the marginal distribution, conditional distribution, and intra-class scatter, while maximizes the inter-class scatter.

Easy Transfer Learning (EasyTL, 2019) [39] which utilizes intra-domain programming to exploit the intra-domain structures.

Discriminative Joint Probability Maximum Mean Discrepancy (DJP-MMD, 2020) [40] which explores the transferability and discriminability of the domains under the independence assumption.

Geometrical Preservation and Distribution Alignment (GPDA, 2021) [21] which jointly utilizes the maximum mean discrepancy (MMD), manifold learning, and scatter preservation to learn discriminative and domain-invariant features.

Progressive Distribution Alignment Based on Label Correction (PDALC, 2021) [25] which adopts label correction to align the distribution shift caused by the target pseudo labels.

Discriminant Geometrical and Statistical Alignment (DGSA, 2022) [24] which adopts density peak landmark selection and manifold learning to mine the potential structural information of the two domains.

Incremental Confidence Samples into Classification (ICSC, 2022) [41] which improves DJP-MMD by progressively labeling and adaptive adjustment strategy. During iterations, the inappropriate estimations of the distributions as well as the pseudo labels are corrected.

C. Experimental Setting

For fair comparisons, the best results from the original papers are cited for comparison. If the results for the datasets

TABLE III
CLASSIFICATION ACCURACIES (%) ON OFFICE+CALTECH10 (SURF)

Source	Target	DICD	EasyTL	DGSA	JPDA	GPDA	PDALC	ICSC	AGE-CS
C	A	47.29	52.61	60.00	47.60	43.70	58.10	55.53	**61.38**
C	W	46.44	53.90	51.90	45.76	42.40	56.60	**59.32**	57.63
C	D	49.68	51.59	49.10	46.50	52.20	52.20	54.14	**58.60**
A	C	42.39	42.30	47.20	40.78	40.80	**47.90**	42.21	44.35
A	W	45.08	43.05	53.50	40.68	41.40	54.60	54.24	**64.75**
A	D	38.85	48.41	49.70	36.94	40.10	44.60	50.96	**57.96**
W	C	33.57	35.35	33.70	34.55	31.90	**39.90**	36.69	35.71
W	A	34.13	38.20	40.40	33.82	35.60	**47.20**	40.08	40.81
W	D	89.81	79.62	89.20	88.54	87.30	**94.30**	75.80	91.08
D	C	34.64	36.06	33.90	34.73	32.50	34.00	34.19	**36.95**
D	A	34.45	38.31	**44.60**	34.66	35.70	42.70	40.50	42.80
D	W	91.19	86.10	87.50	91.19	84.80	91.90	85.42	**93.22**
AVERAGE		48.96	50.46	53.39	47.98	47.37	55.33	52.42	**57.10**

TABLE IV
CLASSIFICATION ACCURACIES (%) ON OFFICE31 (RESNET50)

Source	Target	DICD	EasyTL	DGSA	JPDA	GPDA	PDALC	ICSC	AGE-CS
A	D	82.13	85.10	81.53	82.13	85.80	81.12	87.80	**93.17**
A	W	84.28	83.80	81.51	86.04	87.40	81.51	87.80	**93.08**
D	A	73.48	71.80	70.93	71.07	70.60	71.64	74.10	**75.54**
D	W	99.12	95.10	98.11	97.74	98.40	95.97	95.00	98.74
W	A	71.57	69.60	69.86	68.51	72.80	70.11	74.30	**76.11**
W	D	**99.80**	96.80	**99.80**	99.20	99.40	97.79	97.40	99.20
AVERAGE		85.06	83.70	83.62	84.11	85.73	83.02	86.07	**89.31**

are not available, we grid-search the hyper-parameters listed in the methods and report the best results. For the proposed method, we fix $T = 10$, $\tau = 10^{-3}$, and $\alpha = 5$. Then, we grid-search the regularization parameter λ in [0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10], the dimension d in [10, 20, 30, 40, 50, 60, 70, 80, 90, 100], the neighborhood number k in [8, 10, 16, 32, 64], and the smooth parameter δ in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9].

D. Comparison Experiments

The experimental results are shown in Tables III–VII, where A → C denotes that domain A is transferred to domain C. For easy viewing, the highest accuracies are shown in bold.

Results on Office+Caltech10 (SURF): As shown in Table III, AGE-CS achieves 57.10% classification accuracy and outperforms all the compared methods in average. Compared with PDALC, AGE-CS achieves 1.77% improvement on average. Finally, AGE-CS achieves 64.75% classification accuracy on task A → W which is 10.15% higher than PDALC.

Results on Office31 (ResNet50): The results are shown in Table IV. AGE-CS obtains 89.31% classification accuracy which is 3.24% higher than ICSC. Next, AGE-CS achieves 93.17% and 93.08% classification accuracy on tasks A → D and A → W, which is higher than ICSC by 5.37% and 5.28%, respectively. Four best performances out of six tasks are achieved by AGE-CS.

Results on Office-Home (ResNet50): From Table V, AGE-CS achieves nine best performances with 69.66% classification accuracy. In the experiments, AGE-CS shows its competitiveness with PDALC, and achieves 0.66% improvement compared to PDALC.

Results on ImageCLEF-DA (ResNet50): The results are shown in Table VI. AGE-CS, PDALC, and ISCS achieves 90.60%, 89.79%, and 88.83% classification accuracy, respectively. AGE-CS achieves 0.81% improvement compared to the PDALC and five best performances out of six tasks.

Results on COIL20: As shown in Table VII, AGE-CS achieves 99.38% classification accuracy, while GPDA achieves 96.15% classification accuracy. Compared to PDALC and ICSC, AGE-CS achieves 6.67% and 9.38% average improvement, respectively.

Based on the experimental observations, the following conclusions are given:

1) AGE-CS is effective. In the experiments, AGE-CS outperforms PDALC, ICSC, and GPDA, and achieves the best average performances on the five datasets. The promising results might be attributed to the effectiveness of the proposed adaptive supervision graph embedding method. In other words, AGE-CS appropriately measures the similarity between the samples of the two domains, and reduces the dis-

TABLE V
CLASSIFICATION ACCURACIES (%) ON OFFICE-HOME (RESNET50)

Source	Target	DICD	EasyTL	DGSA	JPDA	GPDA	PDALC	ICSC	AGE-CS
Ar	Cl	53.00	52.80	50.10	46.35	52.90	**54.70**	51.70	54.36
Ar	Pr	73.60	72.10	68.50	60.60	73.40	**76.10**	71.30	75.76
Ar	Re	75.70	75.90	74.20	67.62	77.10	79.50	75.70	**80.22**
Cl	Ar	59.70	55.00	51.20	50.52	52.90	63.20	62.00	**64.85**
Cl	Pr	70.30	65.90	67.50	62.81	66.10	75.40	70.70	**76.77**
Cl	Re	70.60	67.60	67.70	62.59	65.60	75.10	70.70	**76.52**
Pr	Ar	60.90	54.40	54.40	51.79	52.90	63.70	62.40	**65.43**
Pr	Cl	49.40	46.90	46.10	47.72	44.90	52.60	50.00	**53.33**
Pr	Re	77.70	74.70	74.50	72.09	76.10	79.80	76.00	**79.89**
Re	Ar	67.90	63.80	60.80	59.99	65.60	**69.30**	68.20	68.85
Re	Cl	56.20	52.30	51.20	49.99	49.70	56.00	52.40	**56.86**
Re	Pr	79.70	78.00	77.30	74.34	79.20	82.60	79.00	**83.04**
AVERAGE		66.23	63.28	61.96	58.87	63.03	69.00	65.84	**69.66**

TABLE VI
CLASSIFICATION ACCURACIES (%) ON IMAGECLEF-DA (RESNET50)

Source	Target	DICD	EasyTL	DGSA	JPDA	GPDA	PDALC	ICSC	AGE-CS
C	I	90.00	91.50	92.00	88.33	92.33	93.83	91.30	**94.17**
C	P	78.17	77.70	78.00	72.42	78.51	80.71	78.70	**81.39**
I	C	93.33	96.00	95.67	90.83	96.33	96.00	94.70	**96.50**
I	P	80.03	78.70	80.03	75.97	79.53	80.88	**80.90**	80.54
P	C	89.00	95.00	93.67	82.00	91.17	94.33	94.70	**96.33**
P	I	83.50	90.30	92.50	78.83	85.67	93.00	92.70	**94.67**
AVERAGE		85.67	88.20	88.65	81.40	87.26	89.79	88.83	**90.60**

TABLE VII
CLASSIFICATION ACCURACIES (%) ON COIL20

Source	Target	DICD	EasyTL	DGSA	JPDA	GPDA	PDALC	ICSC	AGE-CS
COIL1	COIL2	95.69	80.69	90.97	92.08	96.70	92.64	89.72	**99.58**
COIL2	COIL1	93.33	78.61	93.19	89.86	95.60	92.78	90.28	**99.17**
AVERAGE		94.51	79.65	92.08	90.97	96.15	92.71	90.00	**99.38**

crepancies of the two domains during the iteration. As a result, the appropriate data structure is learned, while the distributions are well-aligned.

2) AGE-CS is stable. In the experiments, some compared methods might lose their competitiveness for some specific datasets. For example, ICSC achieves 86.07% classification accuracy on Office31, while obtaining 90% classification accuracy on COIL. In contrast, AGE-CS achieves all the best performance on the involved datasets, which demonstrates that AGE-CS is considered to be comprehensive.

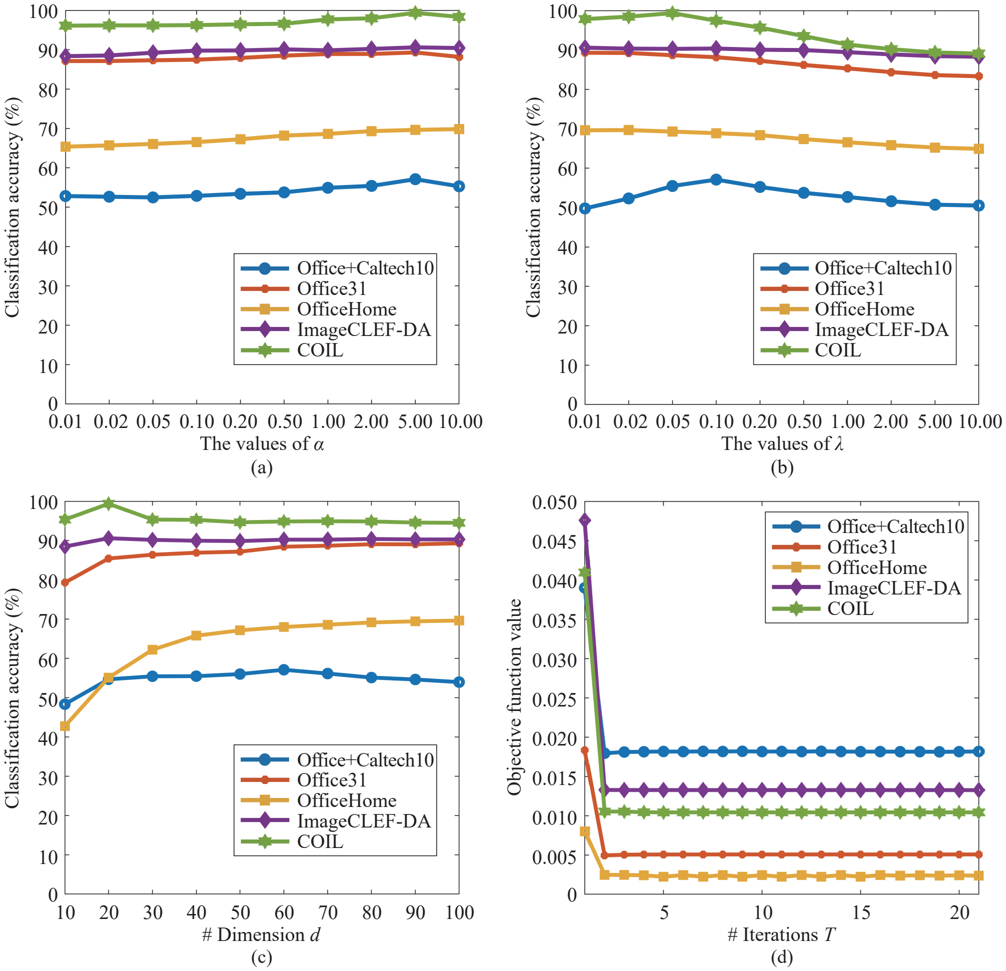
E. Parameter Sensitivity and Convergence Analysis

The sensitivity of parameter α: As shown in Fig. 2(a), the results show that classification accuracy increases with the increase of α ($\alpha \leq 5$), and decreases when $\alpha = 10$. Obviously, an appropriate value of α can facilitate the transfer of the two domains, but a large value of α might lead to large discrepancies. In this case, we propose to set $\alpha = 5$.

The sensitivity of parameter λ: The results are shown in Fig. 2(b). The change in λ has a bit of impact on classification accuracy. AGE-CS achieves the best performances when $\lambda = 0.1$ on the Office+Caltech and $\lambda = 0.05$ on the COIL. For the other datasets, the best performances are achieved when $\lambda = 0.01$. With the above observations, λ can be set as 0.01 for most datasets, and changed for some specific tasks.

The sensitivity of dimension d: The results are shown in Fig. 2(c). AGE-CS achieves the best performance when $d = 20$ on COIL and ImageCLEF-DA, $d = 100$ on Office31 and Office-Home, and $d = 60$ on Office+Caltech10, respectively. Moreover, the performance becomes stable when $d \in [30, 100]$. Therefore, we can fix $d = 100$ for most datasets and change it according to specific tasks.

Convergence Analysis with respect to the number of iteration T: We fix $T = 20$ and run AGE-CS to analyze the



(a) (b) (c) (d)

Fig. 2. The sensitivity of hyper-parameters with respect to α, λ, d, and the convergence analysis of the iteration with respect to T.

convergence of AGE-CS. For the observation purpose, the objective function values are recorded. As shown in Fig. 2(d), the objective function value decreases with the increase in iteration. The results indicate that the proposed method has a good convergence property.

In some emerging fields, there may not be any labels for the target domain, which makes the choice of hyper-parameters more difficult. As an empirical result of the ablation experiment, we let $d = 100$, $\alpha = 5$, $\lambda = 0.1$ (or 0.01), $k = 10$, and $\sigma = 0.1$, and verify the effectiveness of AGE-CS by some parameter-free metrics [42]. Then, a series of heuristic search strategies [43] on λ and σ can be conducted to achieve advanced performance, and be applied in big data scenarios.

F. Ablation Study

The ablation experiments are conducted on the five datasets. We fix GECS as the basic component, and add the combinations of the other two components to it. For simplicity, we denote the components: 1) maximum mean discrepancy (MMD); and 2) AGE. The results are shown in Table VIII, and the ablation methods of AGE-CS are as follows:

- GECS: the method that removes both MMD and AGE, and use GECS only;
- GECS+AGE: the method that uses GECS and AGE;
- GECS+MMD: the method that uses GECS and MMD;
- AGE-CS: the method that uses GECS, MMD, and AGE.

TABLE VIII
THE ABLATION STUDY OF AGE-CS ON FIVE DATASETS

Methods/Accuracy		GECS	GECS+ AGE	GECS+ MMD	AGE-CS
Components	MMD			✓	✓
	AGE		✓		✓
Datasets	Office+Caltech10	51.67	52.08	54.79	**57.10**
	Office31	83.79	83.90	87.89	**89.31**
	Office-Home	67.20	64.01	69.04	**69.66**
	ImageCLEF-DA	86.24	88.09	88.44	**90.60**
	COIL	83.19	82.43	91.88	**99.38**

From the results of Table VIII, we can draw the following conclusions:

a) MMD is important for transfer tasks. In the experiments, the results of GECS, GECS+MMD, and AGE-CS show that MMD is vital to the proper measurement of similarity. If the discrepancies between the two domains are large, AGE might fail to generate a compact similarity matrix. As a result, performance is degraded. In contrast, AGE-CS achieves better performance than GECS+MMD, with 2.31%, 1.42%, 0.62%, 2.16%, and 7.5% improvements on Office+Caltech10, Office31, Office-Home, ImageCLEF-DA, and COIL, resepectively.

b) AGE helps improve performance. The experimental results of GECS and GECS+AGE show that AGE promotes the integration of the two domains, and improves classification performance. Finally, by comparing GECS+AGE and AGE-CS, we find that MMD facilitates AGE to achieve the better performance, which confirms the significance of reducing the domain discrepancies.

V. Conclusion

In this paper, we propose a method called adaptive graph embedding with consistency and specificity (AGE-CS) to address two problems of graph embedding. AGE-CS includes two parts: graph embedding with consistency and specificity (GECS), and adaptive graph embedding (AGE). GECS jointly learns the similarity of samples under the geometric distance and semantic similarity metrics, while AGE adaptively adjust the relative importance of them. By AGE-CS, compact structures are preserved while discrepancies are reduced. Both the experimental results conducted on five datasets and the ablation study verify the effectiveness of AGE-CS.

Limitations: Although AGE-CS achieves promising results, there are three problems that need to be studied in depth:

1) Research on Adaptive Strategies: In this study, we propose Theorem 1 to adaptively tune the hyper-parameter β of semantic graph G. Although some promising results are achieved, the relative importance of distribution alignment and geometric structure is not well addressed. In reality, a parameter-free algorithm is more promising for broad applications. Hence, further studies on the latent relationship between constraints are required.

2) Research on Incomplete Data: Since AGE-CS measures the geometric and semantic distance effectively, it assumes that the data is complete. In reality, there are some domains with incomplete features. In this case, AGE-CS might not work well. A promising way is to complement this incomplete data with information from its nearest neighbors [44], which is left as our follow-up work.

3) Research on Effective Algorithm: Due to the serial nature of generalized eigen-decomposition problem, AGE-CS is difficult to extend as a parallel algorithm. In this case, an application of the gradient descent approach [45] may help AGE-CS solve this tricky problem.

Appendix

Proof of Theorem 1: Equation (5) can be written as

$$\min_{s_{i,:}^T\mathbf{1}=1,0\le s_{i,j}\le 1} \sum_{i,j=1}^{n} d_{i,j}s_{i,j} + \beta_i \sum_{i,j}^{n}(s_{i,j}-g_{i,j})^2 + \gamma(s_{i,j})^2 \tag{29}$$

$\forall i \in [1,n]$, (29) can be further written in a vector form as

$$\min_{s_{i,:}^T\mathbf{1}=1,0\le s_{i,j}\le 1} \left\| s_{i,:} + \frac{1}{2\gamma_i}d_{i,:} \right\|_2^2 + \frac{\beta_i}{\gamma_i}\left\| s_{i,:} - g_{i,:} \right\|_2^2 \tag{30}$$

where γ_i is the optimal parameter of x_i with respect to γ.

By introducing the Lagrangian operator, (30) can be solved by

$$\min_{s_{i,:}} \left\| s_{i,:} + \frac{1}{2\gamma_i}d_{i,:} \right\|_2^2 + \frac{\beta_i}{\gamma_i}\left\| s_{i,:} - g_{i,:} \right\|_2^2 + \xi(s_{i,:}^T\mathbf{1}-1)$$
$$\text{s.t. } 0 \le s_{i,j} \le 1 \tag{31}$$

which indicates that the optimal solution of $s_{i,j}$ is

$$s_{i,j} = \left(\frac{1}{\gamma_i+\beta_i}\left(-\frac{1}{2}d_{i,j}+\beta_i g_{i,j}\right)+\xi\right)_+ \tag{32}$$

where ξ is a constant that makes $s_{i,j} \ge 0$.

For the k-nearest neighbor clustering, we have $s_{i,k+1} \le 0$. Therefore,

$$s_{i,:}^T\mathbf{1} = \sum_{j=1}^{k} \hat{s}_{i,j} = 1. \tag{33}$$

By bringing (32) into (33), we get

$$\sum_{j=1}^{k}\left(\left(-\frac{1}{2}\hat{d}_{i,j}+\beta_i\hat{g}_{i,j}\right)+\xi\right) = 1. \tag{34}$$

Hence, the value of ξ is

$$\begin{aligned}\xi &= \frac{1}{k} - \sum_{j=1}^{k}\hat{s}_{i,j}\\ &= \frac{1}{k} - \frac{1}{\gamma_i+\beta_i}\sum_{j=1}^{k}\left(-\frac{1}{2}\hat{d}_{i,j}+\beta_i\hat{g}_{i,j}\right).\end{aligned} \tag{35}$$

Because $0 \le \hat{s}_{i,j} \le 1$, $\hat{s}_{i,k} > 0$, and $\hat{s}_{i,k+1} \le 0$, the following triangle inequalities should be satisfied:

$$\begin{cases} \dfrac{1}{\gamma_i+\beta_i}\left(-\tfrac{1}{2}\hat{d}_{i,k}+\beta_i\hat{g}_{i,k}\right)+\xi > 0 \\ \dfrac{1}{\gamma_i+\beta_i}\left(-\tfrac{1}{2}\hat{d}_{i,k+1}+\beta_i\hat{g}_{i,k+1}\right)+\xi \le 0. \end{cases} \tag{36}$$

By combining (35) and (36), we get

$$\begin{cases} \dfrac{1}{2}\left(k\hat{d}_{i,k}-\sum_{j=1}^{k}\hat{d}_{i,j}\right)-\beta_i\left(k\hat{g}_{i,k}-\sum_{j=1}^{k}\hat{g}_{i,j}\right)-\beta_i \le \gamma_i \\ \dfrac{1}{2}\left(k\hat{d}_{i,k+1}-\sum_{j=1}^{k}\hat{d}_{i,j}\right)-\beta_i\left(k\hat{g}_{i,k+1}-\sum_{j=1}^{k}\hat{g}_{i,j}\right)-\beta_i > \gamma_i. \end{cases} \tag{37}$$

Let

$$A_i(p) = \frac{1}{2}\left(\hat{d}_{i,p} - \sum_{j=1}^{k}\hat{d}_{i,j}\right) \tag{38}$$

and

$$B_i(p) = -\left(k\hat{g}_{i,p} - \sum_{j=1}^{k}\hat{g}_{ij}\right). \tag{39}$$

Bringing (38) and (39) to (37), we obtain

$$A_i(k)+\beta_i B_i(k)-\beta_i \le \gamma_i < A_i(k+1)+\beta_i B_i(k+1)-\beta_i. \tag{40}$$

Obviously, when RHS[1] of γ_i is greater than LHS[2] of γ_i, the value of γ_i makes sense. Hence, β_i should satisfy

[1] Right hand side

[2] Left hand side

$$\begin{aligned} & A_i(k+1)+\beta_i B_i(k+1) > A_i(k)+\beta_i B_i(k) \\ \Leftrightarrow\quad & (B_i(k+1)-B_i(k))\beta_i > A_i(k)-A_i(k+1) \\ \Leftrightarrow\quad & (-k(\hat{g}_{i,k+1}-\hat{g}_{i,k}))\beta_i > A_i(k)-A_i(k+1). \end{aligned} \tag{41}$$

Since $A_i(k)-A_i(k+1)=\frac{1}{2}\left(\hat{d}_{i,k}-\hat{d}_{i,k+1}\right)\leq 0$, for all $i\in[1,n]$, β_i should satisfy

$$\begin{cases} \beta_i > -\dfrac{A_i(k)-A_i(k+1)}{k(\hat{g}_{i,k+1}-\hat{g}_{i,k})}, & \hat{g}_{i,k} > \hat{g}_{i,k+1} \\ \beta_i \leq -\dfrac{A_i(k)-A_i(k+1)}{k(\hat{g}_{i,k+1}-\hat{g}_{i,k})}, & \hat{g}_{i,k} < \hat{g}_{i,k+1}. \end{cases} \tag{42}$$

Inequality (42) indicates that β_i can be given as

$$\begin{cases} \beta_i = -\dfrac{A_i(k)-A_i(k+1)}{k(\hat{g}_{i,k+1}-\hat{g}_{i,k})}+\epsilon, & \hat{g}_{i,k}-\hat{g}_{i,k+1} > 0 \\ \beta_i = -\dfrac{A_i(k)-A_i(k+1)}{k(\hat{g}_{i,k+1}-\hat{g}_{i,k})}-\epsilon, & \hat{g}_{i,k}-\hat{g}_{i,k+1} < 0 \end{cases} \tag{43}$$

where ϵ is a very small positive number.

Since $\lim_{\epsilon}\epsilon=0$, within the error tolerance, (43) indicates that the value of β_i can be given by

$$\beta_i = \begin{cases} -\dfrac{A_i(k)-A_i(k+1)}{k(\hat{g}_{i,k+1}-\hat{g}_{i,k})} = -\dfrac{\hat{d}_{i,k+1}-\hat{d}_{i,k}}{k(\hat{g}_{i,k+1}-\hat{g}_{i,k})}, & \hat{g}_{i,k+1}\neq\hat{g}_{i,k} \\ \tau, & \hat{g}_{i,k+1}=\hat{g}_{i,k} \end{cases} \tag{44}$$

where $\tau\in\mathbb{R}$ is an arbitrary value used to emphasize semantic information.

When β_i is learned, we can set γ as same as [34]. That is,

$$\begin{aligned} \gamma &= \frac{1}{n}\sum_{i=1}^{n}\gamma_i = \frac{1}{n}\sum_{i=1}^{n}(A_i(k)+\beta_i B_i(k)-\beta_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{2}(k\hat{d}_{i,k}-\sum_{j=1}^{k}\hat{d}_{i,j})-\beta_i(k\hat{g}_{i,k}-\sum_{j=1}^{k}\hat{g}_{i,j}+1)\Big). \end{aligned} \tag{45}$$

■

References

[1] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[2] M. F. Aslan, M. F. Unlersen, K. Sabanci, and A. Durdu, "CNN-based transfer learning–BiLSTM network: A novel approach for COVID-19 infection detection," *Appl. Soft Comput.*, vol. 98, 2021, DOI: 10.1016/j.asoc.2020.106912.

[3] E. F. Ohata, G. M. Bezerra, J. V. S. das Chagas, A. V. L. Neto, A. B. Albuquerque, V. H. C. de Albuquerque, and P. Reboucas Filho, "Automatic detection of COVID-19 infection using chest x-ray images through transfer learning," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 1, pp. 239–248, 2020.

[4] S. Khan, N. Islam, Z. Jan, I. U. Din, and J. J. C. Rodrigues, "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning," *Pattern Recognit. Lett.*, vol. 125, pp. 1–6, 2019.

[5] G. Michau and O. Fink, "Unsupervised transfer learning for anomaly detection: Application to complementary operating condition transfer," *Knowledge-Based Syst.*, vol. 216, 2021, DOI: 10.1016/j.knosys.2021.106816.

[6] X. Wang, X. Liu, and Y. Li, "An incremental model transfer method for complex process fault diagnosis," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 5, pp. 1268–1280, 2019.

[7] S. Teng, N. Wu, H. Zhu, L. Teng, and W. Zhang, "SVM-DT-based adaptive and collaborative intrusion detection," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 1, pp. 108–118, 2018.

[8] Y. Wang, S. Qiu, D. Li, C. Du, B.-L. Lu, and H. He, "Multi-modal domain adaptation variational autoencoder for EEG-based emotion recognition," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 9, pp. 1612–1626, 2022.

[9] H. Hu, H. Wang, Z. Liu, and W. Chen, "Domain-invariant similarity activation map contrastive learning for retrieval-based long-term visual localization," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 2, pp. 313–328, 2022.

[10] L. Zhang and X. Gao, "Transfer Adaptation Learning: A Decade Survey," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022, DOI: 10.1109/TNNLS.2022.3183326.

[11] L. Feng, F. Qian, X. He, Y. Fan, H. Cai, and G. Hu, "Transitive transfer sparse coding for distant domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal*, 2021, pp. 3165–3169.

[12] M. Meng, Y. Liu, and J. Wu, "Robust discriminant projection via joint margin and locality structure preservation," *Neural Proc. Lett.*, vol. 53, no. 2, pp. 959–982, 2021.

[13] E. Gholenji and J. Tahmoresnezhad, "Joint local and statistical discriminant learning via feature alignment," *Signal, Image, Video Proc.*, vol. 14, no. 3, pp. 609–616, 2020.

[14] M. Wang, G. Liu, Zh ao, C. Yan, and C. Jiang, "Behavior consistency computation for workflow nets with unknown correspondence," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 1, pp. 281–291, 2018.

[15] Y. Huang, Z. Shen, F. Cai, T. Li, and F. Lv, "Adaptive graph-based generalized regression model for unsupervised feature selection," *Knowledge-Based Syst.*, vol. 227, 2021, DOI: 10.1016/j.knosys.2021.107156.

[16] B. Han, Q. Yao, T. Liu, G. Niu, I. W. Tsang, J. T. Kwok, and M. Sugiyama, "A survey of label-noise representation learning: Past, present and future," arXiv preprint arXiv: 2011.04406, 2020.

[17] J. Liu, J. Li, and K. Lu, "Coupled local-global adaptation for multi-source transfer learning," *Neurocomputing*, vol. 275, pp. 247–254, 2018.

[18] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P. S. Yu, "Visual domain adaptation with manifold embedded distribution alignment," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 402–410.

[19] S. Vascon, S. Aslan, A. Torcinovich, T. van Laarhoven, E. Marchiori, and M. Pelillo, "Unsupervised domain adaptation using graph transduction games," in *Proc. Int. IEEE Joint Conf. Neural Netw.*, 2019, pp. 1–8.

[20] T. Xiao, L iu, W. Zhao, H. Liu, and X. Tang, "Structure preservation and distribution alignment in discriminative transfer subspace learning," *Neurocomputing*, vol. 337, pp. 218–234, 2019.

[21] 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.

[22] S. Li, S. Song, G. Huang, Z. Ding, and C. Wu, "Domain invariant and class discriminative feature learning for visual domain adaptation," *IEEE Trans. Image Proc.*, vol. 27, no. 9, pp. 4260–4273, 2018.

[23] J. Li, M. Jing, K. Lu, L. Zhu, and H. T. Shen, "Locality preserving joint transfer for domain adaptation," *IEEE Trans. Image Proc.*, vol. 28, no. 12, pp. 6103–6115, 2019.

[24] J. Zhao, L. Li, F. Deng, H. He, and J. Chen, "Discriminant geometrical and statistical alignment with density peaks for domain adaptation," *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 1193–1206, 2022.

[25] Y. Li, D. Li, Y. Lu, C. Gao, W. Wang, and J. Lu, "Progressive distribution alignment based on label correction for unsupervised domain adaptation," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2021, pp. 1–6.

[26] M. Ghifary, D. Balduzzi, W. B. Kleijn, and M. Zhang, "Scatter component analysis: A unified framework for domain adaptation and domain generalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1414–1430, 2016.

[27] Z. Peng, W. Zhang, N. Han, X. Fang, Ka ng, and L. Teng, "Active transfer learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 1022–1036, 2019.

[28] Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey," *Knowledge-Based Syst.*, vol. 151, pp. 78–94,

2018.

[29] S. Rezaei and J. Tahmoresnezhad, "Discriminative and domain invariant subspace alignment for visual tasks," *Iran J. Comput. Sci.*, vol. 2, no. 4, pp. 219–230, 2019.

[30] S. Noori Saray and J. Tahmoresnezhad, "Joint distinct subspace learning and unsupervised transfer classification for visual domain adaptation," *Signal*, *Image and Video Proc.*, vol. 15, no. 2, pp. 279–287, 2021.

[31] M. Jing, J. Li, K. Lu, J. Liu, and Z. Huang, "Adaptive component embedding for unsupervised domain adaptation," in *Proc. IEEE Int. Conf. Multimedia Expo*., 2019, pp. 1660–1665.

[32] Q. Wang and T. Breckon, "Unsupervised domain adaptation via structured prediction based selective pseudo-labeling," in *Proc. AAAI Conf. Artificial Intell.*, vol. 34, no. 4, 2020, pp. 6243–6250.

[33] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 977–986.

[34] F. Nie, C.-L. Wang, and X. Li, "K-multiple-means: A multiple-means clustering method with specified *k* clusters," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2019, pp. 959–967.

[35] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 2066–2073.

[36] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. European Conf. Comput. Vision*. Springer, 2010, pp. 213–226.

[37] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 5018–5027.

[38] S. A. Nene, S. K. Nayar, H. Murase *et al.*, "Columbia object image library (coil-100)," *Technical report, Columbia University*, 1996.

[39] J. Wang, Y. Chen, H. Yu, M. Huang, and Q. Yang, "Easy transfer learning by exploiting intra-domain structures," in *Proc. IEEE Int. Conf. Multimedia Expo*., 2019, pp. 1210–1215.

[40] W. Zhang and D. Wu, "Discriminative joint probability maximum mean discrepancy (DJP-MMD) for domain adaptation," in *Proc. Int. IEEE Joint Conf. Neural Netw*., 2020, pp. 1–8.

[41] S. Teng, Z. Zheng, N. Wu, L. Fei, and W. Zhang, "Domain adaptation via incremental confidence samples into classification," *Int. J. Intell. Syst.*, vol. 37, no. 1, pp. 365–385, 2022.

[42] L. Hu, X. Yuan, X. Liu, S. Xiong, and X. Luo, "Efficiently detecting protein complexes from protein interaction networks via alternating direction method of multipliers," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 6, pp. 1922–1935, 2019.

[43] X. Luo, Y. Yuan, S. Chen, N. Zeng, and Z. Wang, "Position-transitional particle swarm optimization-incorporated latent factor analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 8, pp. 3958–3970, 2022.

[44] D. Wu, Q. He, X. Luo, M. Shang, Y. He, and G. Wang, "A posterior-neighborhood-regularized latent factor model for highly accurate web service QoS prediction," *IEEE Trans. on Services Comput.*, vol. 15, no. 2, pp. 793–805, 2022.

[45] X. Luo, W. Qin, A. Dong, K. Sedraoui, and M. Zhou, "Efficient and high-quality recommendations via momentum-incorporated parallel stochastic gradient descent-based learning," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 2, pp. 402–411, 2021.

Shaohua Teng (Member, IEEE) received his Ph. D. Degree in Industrial Engineering from Guangdong University of Technology, Guangdong, China, in 2008. He is currently a Professor and Dean at the Department of Artificial Intelligent and Information Engineering, School of Advanced Manufacturing, Guangdong University of Technology, Guangzhou, China. He is a member of Association for Computing Machinery (ACM), a distinguished member of China Computer Federation (CCF), and a senior member of Chinese Association of Automation (CAA). His research interests include Pattern Recognition, Network Security, Collaborative Computing, Artificial Intelligence and its Applications. He is the author of 15 patents on his invention and 300+ journal papers. He has earned Guangdong Excellent Teacher Award and two Provincial Science and Technology Awards. He is a Famous Teacher of Guangdong University of Technology. Dr. Teng is a reviewer of IEEE Transactions on Systems, Man, & Cybernetics: Systems, IEEE/CAA Journal of Automatica Sinica, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Computational Social Systems, and Information Sciences.

Zefeng Zheng received the B.S. degree in software engineering from the College of Mathematics and Computer, Guangdong Ocean University, in 2020, and the M.S. degree from the School of Computer Science and Technology, Guangdong University of Technology, in 2023. He is currently pursuing the Ph.D. degree in computer science and technology from the School of Computer Science and Technology, Guangdong University of Technology. His current research interests include machine learning, clustering, and transfer learning. He has served as a Reviewer for a number of journals.

NaiQi Wu (Fellow, IEEE) received the B.S. degree in electrical engineering from Anhui University of Technology, in 1982, the M.S. and Ph.D. degrees in systems engineering both from Xi'an Jiaotong University, in 1985 and 1988, respectively. From 1988 to 1995, he was with the Shenyang Institute of Automation, Chinese Academy of Sciences, and from 1995 to 1998, with Shantou University. He moved to Guangdong University of Technology in 1998. He joined Macau University of Science and Technology, Taipa, Macao, China, in 2013. He was a Visiting Professor at Arizona State University, USA, in 1999; New Jersey Institute of Technology, USA, in 2004; University of Technology of Troyes, France, from 2007 to 2009; and Evry University, France, from 2010 to 2011. He is currently a Chair Professor at the Department of Engineering Science and Macao Institute of Systems Engineering, Macau University of Science and Technology, Taipa, Macao, China. His research interests include production planning and scheduling, manufacturing system modeling and control, discrete event systems, Petri net theory and applications, intelligent transportation systems, and energy systems. He is the author or coauthor of one book, five book chapters, and 200+ journal papers. Dr. Wu was an Associate Editor of the *IEEE Transactions on Systems, Man, & Cybernetics, Part C*, *IEEE Transactions on Automation Science and Engineering*, *IEEE Transactions on Systems, Man, & Cybernetics: Systems*, and Editor in Chief of *Industrial Engineering Journal*, and is an Associate Editor of *Information Sciences*.

Luyao Teng received the B.S. degree from Monash University, Australia, in 2012, the M.S. degree from University of Melbourne, Australia, in 2014, and the Ph.D. degree from Victoria University, Australia in 2019. She is currently with the School of Information Engineering, Guangzhou Panyu Polytechnic and also with Faculty of Information Technology, Monash University. Her current research interests include pattern recognition and machine learning.

Wei Zhang received the M.S. degree in software engineering from the South China University of Technology, in 2005, and is currently a Professor at the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China. She is a Senior Member of China Computer Federation (CCF). Her research interests include pattern recognition, collaborative computing, artificial intelligence and its applications. She has applied for 11 patents on her invention. She has published 20 papers in international journals. She earned the Provincial Science and Technology Award in 2020.