Attention Regularized Laplace Graph for Domain Adaptation

Lingkun Luo, Liming Chen∗, Senior Member, IEEE, and Shiqiang Hu∗

Abstract—In leveraging manifold learning in domain adaptation (DA), graph embedding-based DA methods have shown their effectiveness in preserving data manifold through the Laplace graph. However, current graph embedding DA methods suffer from two issues: 1) they are only concerned with preservation of the underlying data structures in the embedding and ignore sub-domain adaptation, which requires taking into account intra-class similarity and inter-class dissimilarity, thereby leading to negative transfer; 2) manifold learning is proposed across different feature/label spaces separately, thereby hindering unified comprehensive manifold learning. In this paper, starting from our previous DGA-DA, we propose a novel DA method, namely Attention Regularized Laplace Graph-based Domain Adaptation (ARG-DA), to remedy the aforementioned issues. Specifically, by weighting the importance across different sub-domain adaptation tasks, we propose the Attention Regularized Laplace Graph for class aware DA, thereby generating the attention regularized DA. Furthermore, using a specifically designed FEEL strategy, our approach dynamically unifies alignment of the manifold structures across different feature/label spaces, thus leading to comprehensive manifold learning. Comprehensive experiments are carried out to verify the effectiveness of the proposed DA method, which consistently outperforms the state of the art DA methods on 7 standard DA benchmarks, i.e., 37 cross-domain image classification tasks including object, face, and digit images. An in-depth analysis of the proposed DA method is also discussed, including sensitivity, convergence, and robustness.

Index Terms—Domain adaptation (DA), Attention Regularized Laplace Graph (ARG), manifold learning.

I. INTRODUCTION

SUPERVISED learning was at the forefront of increasing real-world applications within the big data era [52], [59], [61], [69]. However, its success relied greatly on the availability of large amounts of high-quality labeled data and assumed that training and testing data share the same distribution for the effectiveness of the learned model. Unfortunately, in real-life applications, data distribution shifts, due to factors as diverse as environment variations, e.g., lighting, temperature, differences in data capture devices, etc., frequently (though not always) occur. It is thus of capital importance to guarantee that a learned model using a given set of training data, namely the source domain, remains effective on a set of testing data, namely the target domain, despite data distribution shifts. This is precisely the objective of the research field of domain adaptation (DA), which aims to develop theories and techniques for effective learning algorithms, while making use of previously labeled data (source domain) and unlabeled data in the current task (target domain), despite the distribution divergence.

While DA can be semi-supervised by assuming a certain amount of labeled data are available in the target domain, in this paper we are interested in unsupervised DA where we assume that the target domain has no labels. While there also exists an increasing number of deep learning-based unsupervised DA methods, here we focus on shallow DA methods as they are easier to train and can provide insights into the design decisions of deep DA methods. The relationships between shallow and deep DA methods will be further discussed in depth in Sect.II on related works.

State-of-the-art shallow DA methods can be categorized into instance-based [12], [59], feature-based [48], [60], [85], or classifier-based. Classifier-based DA is widely applied in semi-supervised DA as it aims to fit a classifier trained on the source domain data to target domain data through adaptation of its parameters, thereby requiring some labels in the target domain [74]. The instance-based approach generally assumes that 1) the conditional distributions of the source and target domains are identical [88], and 2) a certain portion of the data in the source domain can be reused [59] for learning in the target domain through re-weighting. Feature-based adaptation [16], [24], [53], [72], [75] relaxes such a strict assumption and only requires that there exists a mapping from the input data space to a latent shared feature representation space. This latent shared feature space captures the information necessary for training classifiers for both source and target tasks. In this paper, we propose a novel feature-based DA method, which searches a latent feature space using the Laplace graph, but which is regularized by class aware attention along with a unified manifold alignment framework for an effective DA.
A common method for approaching feature adaptation is to seek a shared latent subspace between the source and target domains via optimization [60], [61]. State-of-the-art features three main lines of approach, namely, data geometric structure alignment-based (DGSA), data distribution centered (DDC) or their hybridization. DDC methods [43], [50], [51], [53] aim to search a latent subspace where the discrepancy between the source and target data distributions is minimized, via various distances, e.g., Bregman divergence [71], Geodesic distance [18], Wasserstein distance [7], [8] or Maximum Mean Discrepancy [22] (MMD). DGSA-based methods seek to align the underlying data geometric structures and can be further categorized into subspace alignment-based DA (DGSA-SA DA) and graph embedding-based DA (DGSA-GE DA). DGSA-SA methods, e.g., [70], [72], [85], aim to align different feature spaces under the reconstruction framework using the low rank constraint and/or sparse representation, while DGSA-GE methods explicitly model the underlying data geometric structure by using Laplace graph representation-based manifold learning, thereby seeking a subspace where the source and target data can be well aligned and interleaved by preserving inherent hidden geometric data structures. By hybridizing the DGSA-GE and DDC approaches, our proposed method, namely ARG-DA, strives to leverage the merits of simultaneous alignment of both data distributions and the underlying data geometric structures between the source and target domains.

Popular DGSA-GE methods align different feature/label spaces using graph embedding techniques and can be further summarized into two main research lines, namely feature space embedding-based DA (DGSA-GE-FE), e.g., [31], [89], and label space embedding-based DA (DGSA-GE-LE), e.g., [42], [53], [90]. As shown in Fig.1, DGSA-GE-FE methods seek a common feature space between the source and the target domain by preserving the underlying data geometric structure, while DGSA-GE-LE methods enforce the alignment of manifold structures across the feature space and the label space.

To gain insights into both DGSA-GE-FE and DGSA-GE-LE methods, we review the cornerstone theoretical result in DA [2], [32], which estimates an error bound of a learned hypothesis $h$ on a target domain as follows:

$$
e_T(h) \leq e_S(h) + d_H(D_S, D_T) + \min\{\mathcal{E}_D[S(h(S)) - f_T(x)], \mathcal{E}_D[D(S) - f_T(x)]\}$$

where $e_S(h)$ denotes the classification error on the source domain, and $d_H(D_S, D_T)$ measures the $H$-divergence [32] between two distributions $(D_S, D_T)$ of the source and target domain. The last term of Eq.(1) represents the difference in labeling functions across the two domains. Following Eq.(1), both the DGSA-GE-FE and DGSA-GE-LE methods are concerned with manifold structure preservation, which potentially shrinks the divergence between different labeling functions proposed on the two domains, thereby reducing term.3 of Eq.(1). Different from DGSA-GE-FE, DGSA-GE-LE methods further leverage the discriminative power induced by the labels available on the source domain data by aligning the underlying data geometric structure in accordance with the data labels, thereby reducing the classification error on the source domain $e_S(h)$ (term.1). However, despite interesting performance enabled DGSA-GE methods, they still suffer from the following two issues:

Lack of Attention: The attention mechanism refers to pooling a sequence or a set of features with different weights in order to compute proper representation of the whole sequence or set [27], [78], thereby making it possible to deal differently with different data features while aggregating. However, by enforcing the maintenance of the geometric structure of the underlying data manifold while searching a joint feature subspace between the source and target domains, Laplace graph-based embedding DA methods, i.e., DGSA-GE methods, treat equally all the sub-domain adaptations and can lead to cross-domain misalignment as shown in Fig.2. As can be seen, Fig.2.(b,d) depicts the two-dimensional embeddings from their original three-dimensional feature space as shown in Fig.2.(a,c), respectively.

- **P1:** Fig.2.(b) preserves the data geometric structure of the original data space as in Fig.2.(a) but as a result fails to close adaptively the corresponding sub-domains, e.g., ‘1’, since the two sub-domains (‘1’) are distant in the original data manifold as depicted in Fig.2.(a).
- **P2:** Fig.2.(d) succeeds in aligning the data manifold structures with respect to their original data space as in Fig.2.(c), but as a result fails to separate the two sub-domains, namely, ‘1’ and ‘3’, which are geometrically close in the original data space.

The failures of P1 and P2 are due to the fact that traditional graph embedding-based manifold learning only seeks to
preserve the relative distances across different feature spaces. However, solving a DA task requires bringing closer intra-sub-domains across domains, while preserving the relative distances of inter-sub-domains in the novel projected feature space. As a result, in this paper we propose a novel discriminativeness aware graph embedding, namely, Attention Regularized Graph (ARG), to remedy the aforementioned issues as highlighted by P1 and P2.

Incomplete Unification: By aligning the manifold structures across different feature spaces, DGSA-GE-FE methods harness the geometric knowledge transferred across feature spaces, but fail to leverage the discriminative power induced by the labels available in the source domain data. On the other hand, DGSA-GE-LE remedies the previous issue and unifies the manifold structure across the feature space and the label space for label inference, but fails to align the geometric structures across different feature spaces [19] to ensure smooth transfer learning between the source and target domains. Therefore, a hybridization of both DGSA-GE-FE and DGSA-GE-LE can be appealing for comprehensive manifold learning enhanced DA to leverage the advantages of both approaches.

In this paper, we propose a novel DA method, namely, Attention Regularized Laplace Graph-based Domain Adaptation (ARG-DA), to address the aforementioned ‘Lack of Attention’ and ‘Incomplete Unification’ issues. Specifically, the proposed (ARG-DA), 1) optimizes the Attention Regularized Laplace Graph guided by MMD-based distribution measurements over cross-domain sub-tasks, thereby enforcing the attention regularized manifold learning and 2) implements a hybridization strategy, namely, ‘FEEL’, to unify the manifold structure across the feature/label spaces for comprehensive manifold alignment.

To conclude, the contributions of this paper are summarized as follows:

- The Attention Regularized Laplace Graph is designed to enable discriminative manifold embedding, guided by MMD-based distribution measurements over cross-domain sub-tasks, thereby reducing negative transfer.
- A comprehensive manifold unification strategy, namely, FEEL, is proposed for comprehensive manifold learning, which aligns the manifold structure across different feature spaces to minimize term.3 of Eq.(1), while enforcing the alignment across feature space and label space so as to reduce term.1 of Eq.(1).
- We carry out extensive experiments on 37 image classification DA tasks through 7 popular DA benchmarks and verify the effectiveness of the proposed method, which consistently outperforms thirty-three state-of-the-art DA algorithms by a significant margin. Moreover, we also carry out an in-depth analysis of the proposed DA methods, in particular, w.r.t. their hyper-parameters and convergence speed. In addition, using real data, we also provide insights into the proposed DA model by analyzing the robustness of our proposed ARG term and FEEL strategy.

The paper is organized as follows. Sect.II discusses the related work. Sect.III presents the method. Sect.IV benchmarks the proposed DA method and provides an in-depth analysis. Sect.V draws the conclusion.

II. RELATED WORK

State-of-the-art DA techniques feature two main research lines: 1) Shallow DA; 2) Deep DA. These are overviewed in Sect.II-A and Sect.II-B, respectively, and discussed in comparison with the proposed ARG-DA simultaneously.

A. Shallow Domain Adaptation

1) Geometric Alignment-Based DA: The rationale of geometric alignment-based DA argues that the domain divergence can be reduced by aligning the hidden geometric structure across different domains, while the theoretical explanation is to minimize term.3 of Eq.(1). Recent research in geometric alignment-based DA can be distinguished based on whether it incorporates the graph embedding techniques or embraces the subspace alignment approaches.

Graph Embedding-based Domain Adaptation (DGSA-GE):

By using the graph embedding techniques, DGSA-GE explicitly aligns the hidden manifold across different feature/label spaces for better function learning. To be specific, DGSA-GE can be categorized into DGSA-GE-FE and DGSA-GE-LE groups by aligning the manifold structure among different feature/label spaces.

(a). Feature Space Embedding-based DA (DGSA-GE-FE): DGSA-GE-FE methods [31], [89] aim to search the newly projected common feature space for cross-domain alignment through aligning the manifold structure in the original feature
space. For instance, ARTL [47] uses the GE strategy to align the different feature spaces and to optimize the MMD distance in the unified framework. Jingjing Li et al. [38], [39] proposes cosine similarity to weight the diversity among different samples for local consistency preservation-based DA. To avoid the geometric structure distortion that existed in the original space, MEDA [81] specifically learns the Grassmann manifold to remedy the raised issue. Different from traditional methods, ACE [31] argues that the previous research significantly ignores the discriminative effectiveness, thus, proposes to improve the discriminative effectiveness of the graph model by increasing the within-class distance and reducing the between-class distance across different sub-domains.

LGA-DA [91] assumes that the source and target domains can share the same separation structure and ensures effective DA, by aligning their corresponding Laplace graphs’ eigenspaces. In RHGM [9], hyper-graphs are used to find matches between samples of the source and target domains using similarities of different orders between graphs and achieve a hyper-graph matching based DA.

However, despite huge progress achieved by DGSA-GE-FE methods, DGSA-GE-FE methods only make use of the manifold alignment across different feature spaces to reduce term.3 of Eq.(1), which fails to uniformly align the label space for a comprehensive manifold unification-based DA, thereby preventing it from harnessing the discriminative effectiveness of label space for shrinking of term.1 in Eq.(1).

(b), Label Space Embedding-based DA (DGSA-GE-LE):

Different from DGSA-GE-FE, DGSA-GE-LE explores better usage of the discriminative label space for proper functional learning on the source domain, thereby additionally reducing term.1 of Eq.(1). For this purpose, EDA [90] and DTLC [42] enforce the manifold structure alignment across the feature space and the label space, thereby generating the discriminative conditional distribution adaptation. By incorporating the label smoothness consistency (LSmC), UDA-LSC [26] well accounts for the geometric structures of the underlying data manifold of the label space on the target domain, while DGA-DA [53] further improves UDA-LSC by additionally dealing with manifold regularization on the source domain simultaneously. DAS-GA [63] unifies the cross-domain graph bases to align the feature space and the label space for a spectrum transfer guided DA.

Unfortunately, DGSA-GE-LE ignores the unification across different feature spaces as proposed in DGSA-GE-FE, thereby sacrificing smooth adaptation and resulting in incomplete manifold unification. More importantly, both the DGSA-GE-LE and DGSA-GE-FE methods fail to weigh the importance across the different sub-tasks corresponding to different sub-domains for Attention Regularized DA, thereby increasing the risk of negative transfer.

Subspace Alignment-based Domain Adaptation (DGSA-SA): Apart from DGSA-GE, an increasing number of DA methods, e.g., [11], [14], [54], [70], [72], [85], emphasize the importance of aligning the underlying data subspace rather than the graph model between the source and target domains for effective DA. In these methods, low-rank and sparse constraints are introduced into DA to extract a low-dimension feature subspace where target samples can be sparsely reconstructed from source samples [70], or interleaved by source samples [85], thereby aligning the geometric structures of the underlying data manifolds. A few recent DA methods, e.g., RSA-CDDA [54], JGSA [88], further propose unified frameworks to reduce the shift between domains both statistically and geometrically. HCA [45] improves JGSA using a homologous constraint on the two transformations for the source and target domains, respectively, to make the transformed domains related and hence alleviate negative domain adaptation. Despite the enormous progress achieved by DGSA-SA, we can see that merely aligning the data subspace across domains is also unable to ensure the theoretical effectiveness of DGSA-SA for shrinking the data distributions as required in term.2 of Eq.(1).

Our proposed ARG-DA improves both DGSA-GE and DGSA-SA by jointly aligning the data distributions discriminatively and embracing the attention mechanism for Attention Regularized manifold learning-based DA. Moreover, ARG-DA embraces the proposed FEEL strategy to align the manifold structure across the whole feature/label spaces for the comprehensive manifold unification aware DA.

2) Statistic Alignment-Based DA (STA-DA): The rationale of STA-DA is to assume that the existing domain divergence among the source and target domains can be significantly reduced by searching the newly optimized common feature space using the statistic measurements. Therefore, the learned knowledge from the source domain can be seamlessly applied to the target domain in the optimized common feature space.

For this purpose, recent research embraces a series of statistic measurements, e.g., Bregman Divergence [71], Wasserstein distance [7], [8], and Maximum Mean Discrepancy (MMD) measurement [58], to enforce the cross domain divergence reduced common feature space, thus subsequently reducing term.2 of Eq.(1). Specifically, by embracing the dimensionality reduction techniques, TCA [58] explicitly minimizes the mismatch between source and target in terms of marginal distribution [58], [71]. JDA [48] builds upon TCA’s framework and further exploits the conditional distribution alignment for effective sub-domain adaptation. Apart from previous STA-DA methods, ILS [24] learns the discriminative latent space using Mahalanobis metric to match statistical properties across different domains. Meanwhile, [51] also explores the discriminative effectiveness as in DA by borrowing the merits of linear discriminant analysis (LDA) to optimize the common feature space.

However, despite this, great success has been achieved by STA-DA methods in reducing term.2 of Eq.(1), which significantly fails to use manifold regularization for additionally dealing with term.3 of Eq.(1). Our proposed ARG-DA hybridizes the effectiveness of both DGSA-GE and STA-DA, and aims to develop an effective DA method to account for both the statistic and geometric aspects.
B. Deep Domain Adaptation

Recently, deep learning techniques shed new light on UDA by borrowing the merits of high discriminative feature representations enabled by deep learning. They featured the following two main approaches.

1) Statistic Matching-Based DA: These approaches share similar spirits to STA-DA by reducing the domain shift via different statistic measurements. However, this domain shift reduction is achieved by incorporating the DL paradigm rather than the shallow functional learning as proposed in STA-DA. The popular DAN [46] reduces the marginal distribution divergence by incorporating the multi-kernel MMD loss on the fully connected layers of AlexNet. JAN [50] improves DAN by jointly decreasing the divergence of both the marginal and conditional distributions. D-CORAL [73] further introduces the second-order statistics into the AlexNet [34] framework for a more effective DA strategy. Different from previous research, FDA [86] and MSDA [23] provide novel insight into shrinking the domain shift. FDA argues that the dissimilarities across domains can be measured by the low-frequency spectrum, while MSDA insists on aligning different domains by narrowing the LAB feature space representation, thereby swapping the contents of different low-frequency spectrum [86], while LAB feature space representation [23] among cross-domains enables qualified domain shift reduction. ATM [36] introduces an original distance loss, namely, maximum density divergence (MDD) to quantify distribution divergence for effective DA, and combines adversarial deep learning with metric learning. BOS [30] projects the cross-domain samples onto a hyper-spherical latent space to implement the angular distance measured distribution alignment within the open set domain adaptation experimental scenarios. CDCL [64] proposes a generic cross-domain collaborative learning framework to collaboratively learn a shared feature representation for an effective cross-domain data analysis. ETD [40] designed an enhanced transport distance to quantify the divergence between the cross-domain samples for attention aware DA. Both SHOT [44] and MA-UDA [41] aim to solve a novel DA setting, namely, source-free domain adaptation (SFDA). SHOT hybridizes the mutual information and self-supervised pseudo-labeling techniques for effective SFDA, while MA-UDA develops a novel generative model to produce target-style training samples for SFDA.

2) Adversarial Loss-Based DA: These methods make use of GAN [21] and propose to align data distributions across domains by making sample features indistinguishable w.r.t the domain labels through an adversarial loss on a domain classifier [16], [62], [75]. DANN [16] and ADDA [75] learn a domain-invariant feature subspace by reducing the marginal distribution divergence. MADA [62] additionally makes use of multiple domain discriminators, thereby aligning conditional data distributions. Different from the previous approaches, DSN [4] achieves domain-invariant representations by explicitly separating the similarities and dissimilarities in the source and target domains. MADAN [92] explores knowledge from different multi-source domains to fulfill DA tasks. CyCADA [25] addresses the distribution divergence using a bi-directional GAN-based training framework. UDA-GCN [83] facilitates knowledge transfer across domains based on the specifically designed dual graph convolutional network. DAGNN [84] designs the domain-adversarial graph neural network to comprehensively capture the semantic knowledge for effective text classification. MDDA [87] hybridizes the multimodal disentangled representation learning and UDA techniques within a unified framework for prompt rumor detection of the newly emerged events.

The main advantage of these DL-based DA methods is that they jointly shrink the divergence of data distributions across domains and achieve a discriminative feature representation of data through a single unified end-to-end learning framework. Therefore, the optimized model naturally benefits from the merits similar to those in META learning [82], which argues that the model can be reinforced by harnessing different tasks so as to receive the best candidate hyper-parameters. However, they generally suffer from the ‘batch learning’ [20] strategy in contrast to ‘global modeling’ [1] based optimization. As a result, the manifold alignment of DL-based DA is generally implemented locally within the randomly selected batch samples rather than the global view enforced graph model [93]. In this research work, our proposed ARG-DA is global manifold structure alignment-based optimization, while the rationale can also be easily explained, thereby enabling insights into further improving DA methods.

III. THE PROPOSED METHOD

In Sect.III, we first define the notations. We set out the DA problem in Sect.III-A, after which Sect.III-B formulates our DA model. Sect.III-C presents the method for solving the proposed DA model and derives the algorithm of ARG-DA. Sect.III-D extends ARG-DA to non-linear problems through kernel mapping.

A. Notations and Problem Statement

Matrices are written as boldface uppercase letters. Vectors are written as boldface lowercase letters. For matrix $M = (m_{ij})$, its $i$-th row is denoted by $\mathbf{m}_i$, and its $j$-th column by $\mathbf{m}_j$. We define the Frobenius norm $\| \cdot \|_F$ as: $\| M \|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} m_{ij}^2}$. A domain $D$ is defined as an $l$-dimensional feature space $x$ and a marginal probability distribution $P(x)$, i.e., $D = \{\chi, P(x)\}$ with $x \in \chi$. Given a specific domain $D$, a task $T$ is composed of a $C$-cardinality label set $\mathcal{Y}$ and a classifier $f(x)$, i.e., $T = \{\mathcal{Y}, f(x)\}$, where $f(x) = \mathbb{Q}(y|x)$ can be interpreted as the class conditional probability distribution for each input sample $x$.

In unsupervised domain adaptation, we are given a source domain $\mathcal{D}_S = \{x_{i}^{l}, y_{i}^{l}\}_{i=1}^{n_l}$ with $n_l$ labeled samples $X_S = [x_{1}^{l} \ldots x_{n_l}^{l}]$, which are associated with their class labels $Y_S = \{y_{1}^{l}, \ldots, y_{n_l}^{l}\} \in \mathbb{R}^{n_l \times C}$, and an unlabeled target domain $\mathcal{D}_T = \{x_{i}^{t}\}_{i=1}^{n_t}$ with $n_t$ unlabeled samples $X_T = [x_{1}^{t} \ldots x_{n_t}^{t}]$, whose labels $Y_T = \{y_{n_l+1}^{t}, \ldots, y_{n_l+n_t}\} \in \mathbb{R}^{n_t \times C}$ are unknown. Here, $y_{i}^{t} \in \{0, 1\}$ if $x_{i}^{t}$ belongs to the $j$-th class, and 0 otherwise. We define the data matrix $X = [X_S, X_T] \in \mathbb{R}^{l \times n}$ ($l$ = feature dimension; $n = n_s + n_t$) in packing both the
source and target data. The source domain $D_S$ and target domain $D_T$ are assumed to be different, i.e., $\mathcal{X}_S \neq \mathcal{X}_T$, $\mathcal{Y}_S = \mathcal{Y}_T$. $P(\mathcal{X}_S) \neq P(\mathcal{X}_T)$, $Q(\mathcal{Y}_S|\mathcal{X}_S) \neq Q(\mathcal{Y}_T|\mathcal{X}_T)$. We also define the notion of sub-domain, i.e., class, denoted as $D^{(c)}_S$, representing the set of samples in $D_S$ with the class label $c$. It is worth noting that the definition of sub-domains in the target domain, namely $D^{(c)}_T$, requires a base classifier, e.g., Nearest Neighbor (NN), to attribute pseudo labels for the samples in $D_T$.

Maximum Mean Discrepancy (MMD) is an effective non-parametric distance-measurement that compares the distributions of two datasets by mapping the data into Reproducing Kernel Hilbert Space [3] (RKHS). Given two distributions $P$ and $Q$, the MMD between $P$ and $Q$ is defined as:

$$\text{Dist}(P, Q) = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(x_i) - \frac{1}{n_2} \sum_{i=1}^{n_2} \phi(x_i) \right\|_{\mathcal{H}}$$

(2)

where $P = \{p_1, \ldots, p_{n_2}\}$ and $Q = \{q_1, \ldots, q_{n_2}\}$ are two random variable sets from distributions $P$ and $Q$, respectively, and $\mathcal{H}$ is a universal RKHS with the Reproducing Kernel Mapping $\phi$: $f(x) = \langle \phi(x), f \rangle$, $\phi: \mathcal{X} \rightarrow \mathcal{H}$.

B. Formulation

As shown in Fig.3, our proposed model (Sect.III-B.4) starts from Close yet Discriminative Domain Adaptation (CDDA) [53], [55] (Sect.III-B.1) to seek a common feature space in which both term.1 and term.2 of Eq.(1) are minimized. Then, we propose in Sect.III-B.2 the Attention Regularized Laplace Graph (ARG)-based DA, guided by similarity and discriminativeness when adapting each sub-domain in both the source and target domains, thereby improving the traditional Laplace graph. Sect.III-B.3 implements the proposed ‘FEEL’ strategy and makes use of the proposed ARG to align the manifold structures across different feature/label spaces for comprehensive manifold unification, thus reducing term.1 and term.3 of Eq.(1) simultaneously. By integrating CDDA and ARG-based manifold learning, Sect.III-B.4 formalizes our final model (ARG-DA) and thereby optimizes at the same time the three terms of the right-hand of Eq.(1).

1) Discriminative Statistic Alignment: Given the fact that the main issue significantly affecting model transferability is the cross-domain distribution divergence, a discriminative approach, i.e., CDDA [53], [55], is first implemented here to reduce the mismatch of the marginal distribution and the conditional distribution between the source and target domains.

a) Matching marginal and conditional distributions: As shown in Fig.4, our model starts from JDA [48], which makes use of MMD in RKHS to measure the distances between the expectations of the source domain/sub-domain and target domain/sub-domain. Specifically, 1) the empirical distance of the source and target domains is defined as $\text{Dist}^m$; 2) the conditional distance $\text{Dist}^c$ is defined as the sum of the empirical distances between pairs of sub-domains in $D_S$ and $D_T$ with the same label; 3) $\text{Dist}_{\text{clo}}$ is defined as the sum of $\text{Dist}^m$ and $\text{Dist}^c$, which will be optimized to bring closer both data distributions and class conditional data distributions across domains.

$$\text{Dist}_{\text{clo}} = \text{Dist}^m(D_S, D_T) + \text{Dist}^c \sum_{c=1}^{C} (\text{Dist}^c(D_S^c, D_T^c))$$

$$= \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} A^T x_i - \frac{1}{n_1} \sum_{i=n_1+1}^{n_2} A^T x_i \right\|^2 + \sum_{c=1}^{C} \left\| \frac{1}{n_c} \sum_{x_i \in D_c^{(c)}} A^T x_i - \frac{1}{n_c} \sum_{x_j \in D_c^{(c)}} A^T x_j \right\|^2$$

(3)

- $\text{Dist}^m(D_S, D_T)$: where $M_0$ is the MMD matrix between $D_S$ and $D_T$ with $(M_0)_{ij} = \frac{1}{n_i n_j}$ if $(x_i, x_j) \in D_S$, $(M_0)_{ij} = \frac{1}{n_i n_j}$ if $(x_i, x_j) \in D_T$ and $(M_0)_{ij} = \frac{1}{n_i n_j}$ otherwise. Thus, the difference between the marginal distributions $P(x_S)$ and $P(x_T)$ is reduced when minimizing $\text{Dist}^m(D_S, D_T)$.

- $\text{Dist}^c(D_S, D_T)$: which is the number of classes, $D(S)^c = \{x_i : x_i \in D_S \land y(x_i) = c\}$ represents the $c^{th}$ sub-domain in the source domain, in which $n_i^c = \| D(S)^c \|$ is the number of samples in the $c^{th}$ source sub-domain. $D_T^c$ and $n_i^c$ are defined similarly for the target domain but using pseudo labels. Finally, $M_c$ denotes the MMD matrix between the sub-domains with labels $c$ in $D_S$ and $D_T$ with $(M_c)_{ij} = \frac{1}{n_i^c n_j^c}$ if $(x_i, x_j) \in D(S)^c$, $(M_c)_{ij} = \frac{1}{n_i^c n_j^c}$ if $(x_i, x_j) \in D(T)^c$, $(M_c)_{ij} = \frac{1}{n_i^c n_j^c}$ if $(x_i \in D(S)^c, x_j \in D(T)^c) \text{ or } (x_i \in D(T)^c, x_j \in D(S)^c)$ and $(M_c)_{ij} = 0$ otherwise. Consequently, the mismatch of conditional distributions between $D_S^c$ and $D_T^c$ is reduced by minimizing $\text{Dist}^c$.

- Finally, the original data $X$ are projected into the optimal common feature space using the mapping $A$ through $A^T X$.

b) Repulsive force (RF) across/within domains: As shown in Fig.4, JDA is merely concerned with shrinking the MMD distances in order to reduce the cross-domain
distribution divergence and ignores discriminative knowledge within data. Therefore, we directly borrow the idea from CDDA and DGA-DA [53], [55] and introduce the Repulsive Force (RF) term \( Dist^{re} = Dist^{re}_{S\rightarrow T} + Dist^{re}_{S\rightarrow S} \) to further explore the discriminativeness of data. While JDA and DGA-DA have so far endeavored to minimize term.2 of Eq.(1), in this paper we additionally optimize term.1 of Eq.(1) as shown in Fig.4 and introduce a novel RF term \( Dist^{re}_{S\rightarrow S} \) as in Fig.4, so as to increase the discriminative power of the labeled source domain data, thereby allowing a better predictive model on the source domain and a decreasing term.1 of Eq.(1). Subsequently, we use \( S \rightarrow T \), \( T \rightarrow S \) and \( S \rightarrow S \) to index the distances computed from \( D_S \) to \( D_T \), \( D_T \) to \( D_S \) and \( D_S \) to \( D_S \), respectively, and \( Dist^{re}_{S\rightarrow T} \) as the sum of the distances between each source sub-domain \( D^{c}_S \) and all the target sub-domains \( D^{(r)}_T \); \( r \in \{[1 \ldots C] \} \) excluding the \( c \)-th target sub-domain. Symmetrically, \( Dist^{re}_{T\rightarrow S} \) and \( Dist^{re}_{S\rightarrow S} \) are defined similarly to \( Dist^{re}_{S\rightarrow T} \). Consequently, the final Repulsive Force term \( Dist^{re} \) is formalized as:

\[
\begin{align*}
Dist^{re}_{S\rightarrow T} + Dist^{re}_{T\rightarrow S} + Dist^{re}_{S\rightarrow S} &= Dist^{r} \sum_{c=1}^{C} (D^{c}_S, D^{r}_{S \in \{1\ldots C\} \setminus \{c\}}) \\
&+ Dist^{r} \sum_{c=1}^{C} (D^{T}_{c}, D^{r}_{S \in \{1\ldots C\} \setminus \{c\}}) \\
&+ Dist^{r} \sum_{c=1}^{C} (D^{c}_S, D^{r}_{S \in \{1\ldots C\} \setminus \{c\}}) \\
&= \sum_{c=1}^{C} \text{tr}(A^T X(M_{S\rightarrow T} + M_{T\rightarrow S} + M_{S\rightarrow S})X^T A) \quad (4)
\end{align*}
\]

where
- \( M_{S\rightarrow T} \) is defined as: \( (M_{S\rightarrow T})_{ij} = \frac{1}{n_i n_j} \) if \( (x_i, x_j) \in D_S^{(c)} \), \( \frac{-1}{n_i n_j} \) if \( (x_i, x_j) \in D_T^{(r)} \) and \( 0 \) otherwise.
- \( M_{T\rightarrow S} \) is defined as: \( (M_{T\rightarrow S})_{ij} = \frac{1}{n_i n_j} \) if \( (x_i, x_j) \in D_T^{(c)} \), \( \frac{-1}{n_i n_j} \) if \( (x_i, x_j) \in D_S^{(r)} \) and \( 0 \) otherwise.
- \( M_{S\rightarrow S} \) is defined as: \( (M_{S\rightarrow S})_{ij} = \frac{1}{n_i n_j} \) if \( (x_i, x_j) \in D_S^{(c)} \), \( \frac{-1}{n_i n_j} \) if \( (x_i, x_j) \in D_S^{(r)} \) and \( 0 \) otherwise.

Therefore, maximizing \( Dist^{re}_{S\rightarrow T} \) and \( Dist^{re}_{S\rightarrow S} \) increases the distances of each sub-domain with the other remaining sub-domains across domains, i.e., the between-class distances across domains, thereby reducing term.2 of Eq.(1). On the other hand, maximizing \( Dist^{re}_{S\rightarrow S} \) increases the between-class distances in the source domain, thereby optimizing term.1 in Eq.(1) on classification errors on the source domain. In our model, the RF term is proposed within the source domain and across the different domains simultaneously for optimizing the first and second terms of the right-hand in Eq.(1) simultaneously.

2) Attention Regularized Laplace Graph: Although Sect.III-B.1 improves CDDA and DGA-DA to optimize Term.1 and Term.2 of Eq.(1), minimization of Term.3 in Eq.(1) is not addressed and further requires manifold regularization in order to decrease the divergence between the classifiers on the source and target domains. However, as discussed in Sect.I, traditional manifold learning-based DA is not concerned with data class information and suffers from the so-called issue of ‘Lack of Attention’. To remedy this issue, we introduce here an Attention Regularized Laplace Graph (ARG) term to explicitly address the issues of P1 and P2 as highlighted in Sect.I and to achieve discriminativeness aware manifold learning. Specifically, the proposed ARG aims to achieve the following two goals:

- **Goal.1:** To avoid the injudicious separation of data from the same sub-domain as shown in Fig.2,(b), the proposed ARG needs to bring closer those cross sub-domains with the same label but with large domain divergence.
- **Goal.2:** To prevent the brute force data geometric structure alignment as highlighted in Fig.2,(d), the designed ARG needs to enable separation of those differently labeled yet closely aligned sub-domains.

**a) Mathematical formalization:** To meet Goal.1 and Goal.2, the ARG term is mathematically formalized through the following steps:

- **Step.1 - Design of an Affinity Matrix:** Initially, an affinity matrix \( W \) is defined to capture the relative relationship of each sample across the whole database, which is also the fundamental framework of the ARG term. Formally, we denote the pair-wise affinity matrix as:

\[
W_{ij} = \begin{cases} 
\text{sim}(x_i, x_j), & x_i \in N_p(x_j) \text{ or } x_j \in N_p(x_i) \\
0, & \text{otherwise}
\end{cases}
\]

where \( W = [w_{ij}]_{(n+n_l) \times (n+n_l)} \) is a symmetric matrix [57], with \( w_{ij} \) giving the affinity between two
data samples \(x_i\) and \(x_j\) and defined as 
\[
sim(x_i, x_j) = \exp(-\|x_i - x_j\|^2_{2})
\]
if \(i \neq j\) and \(w_{ii} = 0\) otherwise. 
\(x_i \in N_p(x_j)\) denotes \(x_i\) belonging to the neighbors of \(x_j\), while \(x_j \in N_p(x_i)\) is defined similarly but based on \(x_i\)'s neighbors. Similar to DGA-DA, the greater the distance between a pair of samples \(dist(x_i, x_j)\), the less their similarity \(\sim(x_i, x_j)\), and vice versa.

- **Step.3 (Intra-Class Attraction Aware Attention)**: We also calculate the conditional distribution divergences, thereby bringing closer these cross sub-domains for class-aware data geometric alignment. For this purpose, as illustrated in Fig.5(c), we define the Intra-Class Attraction Aware Attention (A.Attention) in (7), as shown at the bottom of the next page.

By dot multiplication of Eq.(5) and Eq.(7), the affinity matrix receives an attention regularization, \(W_{A.Attention} = W \bullet A.Attention\), which amplifies the similarity of cross-domain data pairs when they belong to the same sub-domain.

- **Step.4 (RF - Inter-class Conditional Distribution Divergence)**: We also calculate the conditional distribution divergence between differently labeled sub-domains in the source domain using MMD measurement.

\[
RF.Dist^{c,i}_{i \in \{1...C\}, i \neq c}(D_{S}^{c}, D_{S}^{j}) = \left( \frac{1}{n_{c}^{(c)}} \sum_{x_i \in D_{S}^{(c)}} A_{i}^{T} x_i - \frac{1}{n_{j}^{(c)}} \sum_{x_j \in D_{S}^{(j)}} A_{j}^{T} x_j \right)^{2}
\]

(8)

Similar to Eq.(6), Eq.(8) captures the distance between the differently labeled sub-domains in the source domain in terms of conditional distributions. Inspired by the Repulsive Force(RF) term as defined in Eq.(4), we name the distance of Eq.(8) as \(RF.Dist\) to distinguish it from the class attraction distance in Eq.(6).

\[
RF.Dist^{c,i}_{i \in \{1...C\}, i \neq c}(D_{S}^{c}, D_{T}^{j}) = \left( \frac{1}{n_{c}^{(c)}} \sum_{x_i \in D_{S}^{(c)}} A_{i}^{T} x_i - \frac{1}{n_{j}^{(c)}} \sum_{x_j \in D_{T}^{(j)}} A_{j}^{T} x_j \right)^{2}
\]

(9)

As shown in Fig.5(b), we also propose Eq.(9) to capture the distance between the differently labeled sub-domains across the source and target domains.

- **Step.5 (Inter-Class Repulsion Aware Attention Map)**: To meet Goal.2, we need to decrease the similarities of closely confounded yet differently labeled sub-domain data. As shown in Fig.5(d), Inter-Class Repulsion Aware Attention (R.Attention) is defined in (10), as shown at the bottom of the next page.

Using Eq.(10), the affinity matrix can be improved by attention regularization again, using the Inter-Class Repulsion Aware Attention Map (R.AM), through dot multiplication \(W_{R.Attention} = W \bullet R.Attention\), which decreases the affinities of sub-domain data that are differently labeled but very confounded in terms of data distributions.

- **Step.6 (Attention Map)**: As shown in Fig.5(c), by unifying the Intra-Class Attraction Attention Map A.AM and the Inter-Class Repulsion Attention Map R.AM, we obtain our final Attention Map (AM), where the empty position is embedded with ‘1’ to preserve the original manifold structures.

- **Step.7 (Attention Regularized Laplace Graph)**: Finally, the previous Attention Map is used to improve the traditional Laplace graph (L) in order to obtain the Attention Regularized Laplace Graph (ARG), which is mathematically formulated as:

\[
ARG_{AM} = (I - D_{AM}^{-1} W_{AM} D_{AM}^{-1})^{-1} \]

(11)

where \(D_{AM} = \text{diag}(d_{AM,1}, \ldots, d_{AM,n_{(1\ldots1)}})\) is the degree matrix with \(d_{AM,i} = \sum_j w_{AM,i,j}\). The proposed ARG not only enables preservation of data manifold structures by following the standard spectral graph theory, but also modulates it by taking into account intra-class attraction and inter-class repulsion over the source and target domains, thereby leading to discriminative attention aware manifold embedding.

3) FEEL-Based Manifold Alignment: to deal with the second issue (‘Incomplete Unification’) as highlighted in Sect.I, we also propose a novel manifold learning strategy, which hybridizes the DGSA-GE-FL and DGSA-GE-LF approaches into a unified optimization framework and integrates seamlessly ARG-based embeddings, and formulate our final FEEL strategy for dynamic alignment of the manifold structures across different feature/label spaces as illustrated in Fig.6. Specifically, the proposed FEEL strategy aims to achieve the following three goals:

- **Goal.3**: Alignment of the manifold structures between the original feature space and the common feature space as enabled in the DGSA-GE-FL approach, thereby reducing term.3 of Eq.(1).

- **Goal.4**: Alignment of the common feature space and the label space as proposed in the DGSA-GE-LF techniques in order to decrease term.1 of Eq.(1).

- **Goal.5**: Seamless integration of the Attention Regularized Laplace Graph ARG for data distribution modulated graph-based embeddings.
**Implementation details:** To meet *Goal.3* through *Goal.5*, FEEL hybridizes (a) DGSA-GE-FE and (b) DGSA-GE-LE. Furthermore, inspired by DGA-DA [53], we also introduce the (c) Label Smoothness Consistency term (LSmC), to ensure robust functional learning.

(a). **Feature Space Embedding:** as shown in Fig.6.(a), FEEL starts by searching a common feature space (C) from the original feature space (X) using matrix transformation (X → AX), while respecting the underlying data manifold geometric structure, and minimizes the following objective function:

\[
\begin{align*}
\min & \sum_{i,j} (A_{x_i} - A_{x_j})^2 W_{ij}^O \\
= & \min \left( \sum_{i,j} ((A_{x_i})^2 + (A_{x_j})^2 - 2A_{x_i}A_{x_j}W_{ij}^O) \right) \\
= & \min \left( \sum_{i} (A_{x_i})^2 D_{ii}^O + \sum_{j} (A_{x_j})^2 D_{jj}^O - 2 \sum_{i,j} A_{x_i}A_{x_j}W_{ij}^O \right)
\end{align*}
\]

where \(W^O\) denotes the affinity matrix which is based on the Original feature space, and \(D^O = diag(a_{(1,1)}, \ldots, a_{(n_s+n_t), (n_s+n_t)})\) is the degree matrix with \(d_{(i,i)}^O = \sum_j w_{ij}^O\). Optimizing Eq.(12) merely deals with *Goal.3*. Therefore, we propose Eq.(13) to improve Eq.(12) via attention regularization to partially meet *Goal.5*:

\[
\begin{align*}
\min & (A^TXARG^O A^T) \\
s.t. & A^T XARG^O A^T = (I - (D_{AM}^O)^{-\frac{1}{2}} W_{AM}^O (D_{AM}^O)^{-\frac{1}{2}}) \quad (13)
\end{align*}
\]

Eq.(13) differs from Eq.(12) by reformulating \(W^O\) as \(W_{AM}^O\), thereby generating the attention aware DGSA-GE-FE.

(b). **Label Space Embedding:** as shown in Fig.6.(b), in order to meet *Goal.4* and *Goal.5*, we further align the manifold structure across the common feature space (C) and the label space (Y) via ARG\(^C\) regularized functional learning:

\[
\begin{align*}
\min (Y^T ARG^C Y) \\
\text{s.t.} \quad & A^C = (I - (D_{AM}^C)^{-\frac{1}{2}} W_{AM}^C (D_{AM}^C)^{-\frac{1}{2}}) \quad (14)
\end{align*}
\]

in which \(W^C\) is computed in the Common feature space.

(c). **Label Smoothness Consistency:** to enhance the robustness of dynamic functional learning across DGSA-GE-FE and DGSA-GE-LE, we borrow the merits of the label smoothness consistency (LSmC) term [26], [42] as proposed in DGA-DA:

\[
\begin{align*}
\text{Dist}_{\text{label}}^{(i)} = \sum_{j=1}^C \sum_{i=1} \| Y_{i,j} - Y_{i,j}^0 \| \\
\text{where} Y = Y_S \cup Y_T, \quad Y_{i,j} = \arg \max_{c \in Y_S} Y_{i,j}^F \quad (15)
\end{align*}
\]

where \(D_T^{(c)}\) is defined as pseudo labels, generated via a base classifier, e.g., NN. Therefore, jointly optimizing Eq.(13), Eq.(14) and Eq.(15) ensures *Goal.4* through *Goal.5* and guarantees robust optimization simultaneously.

4) **The Final Model:** To sum up, our final model (ARG-DA) jointly optimizes Eq.(3), Eq.(4), Eq.(13), Eq.(14), and Eq.(15) within a unified framework, which comprehensively minimizes the error bound as defined in Eq.(1) and simultaneously addresses the two issues highlighted in Sect.I, namely *(Lack of Attention)* and *(Incomplete Unification)*. Specifically, it can be written mathematically as:

\[
\begin{align*}
\min & \left( \sum_{c=0}^C tr(A^T X M^* + ARG^O A^T A) + \lambda \| A \|_F^2 \right) \\
+ \mu \left( \sum_{j=1}^C \sum_{i=1} \| Y_{i,j}^F - Y_{i,j}^0 \| + Y^T (ARG^C Y) \right)
\end{align*}
\]
problems following three key factors:

C. Solving the Model

Optimization of our final model Eq.(17) is based on the following three key factors:

- **F.1**: Discriminative statistic alignment (Eq.(3), Eq.(4));
- **F.2**: ARG-based DGSA-GE-FE, which aligns the manifold structure across different feature spaces, (Eq.(13));
- **F.3**: ARG-based DGSA-GE-LE, which aligns the manifold structure across feature and label spaces, (Eq.(14), Eq.(15)).

These three key factors can be categorized into two subproblems (SPs) for the dynamic functional learning:

- **SP.1**: in SP.1, both F.1 and F.2 enforce the domain alignment across the original feature space and the common feature space. Mathematically, SP.1 can be formalized as:
  \[
  \min \sum_{c=0}^{C} tr(A^T X (M^c + ARG^O) X^T A) + \lambda \|A\|^2
  \]

- **SP.2**: using the updated feature representation \(A^T X\) of SP.1. SP.2 aligns the manifold structure across the common feature space and the label space to achieve F.3.
  \[
  \min_{\mu, \gamma} \sum_{j=1}^{C} \sum_{i=1}^{n_j \ast n_i} \|Y_{ij}^T - Y_{ij}^0\|^2 + Y^T (ARG^C) Y
  \]

As a result, our final model Eq.(17) can thus be naturally divided into two sub-optimization problems (Eq.(18) and Eq.(19)), each of which has a closed form solution and can be optimized alternatively.

Optimization of **SP.1**: Eq.(18) amounts to solving the generalized eigendecomposition problem in order to obtain the best matrix projection \(A\). By using the Augmented Lagrangian method [15], [48], we obtain the best candidate matrix projection through setting its partial derivation w.r.t. \(A\) equal to zero:

\[
(X(M^c + ARG^O)X^T + \lambda I)A = XHX^TA\Phi
\]

where \(\Phi = \text{diag}(\phi_1, \ldots, \phi_k) \in R^{k \times k}\) is the Lagrange multiplier, and \(ARG^O\) is calculated using the original feature representation. The optimal subspace \(A\) is reduced to solving Eq.(20) for the \(k\) smallest eigenvectors. Then, we obtain the projection matrix \(A\) and the updated feature representation in the newly found common feature space \(Z = A^T X\).

Optimization of **SP.2**: using \(Z\) obtained in SP.1, we formulate \(ARG^C\) to further explore the discriminative power of the label space \(Y\) by optimizing Eq.(19). Inspired by previous research [33], [93], the minimum of Eq.(19) is approached where the derivative of the function is zero. Therefore, the approximate solution can be formulated as:

\[
Y^* = (D_{AM}^C - (1/\gamma)W_{AM}^C)^{-1}Y^{(0)}
\]

where \(Y^*\) is the probability of prediction of the target domain corresponding to different class labels, \(W_{AM}^C\) is the affinity matrix based on the feature representation in the optimized common feature space, and \(D_{AM}^C\) is the diagonal matrix.

For simplicity’s sake, we define \(\alpha = 1/\gamma\), then Eq.(21) is reformulated as Eq.(22):

\[
Y^* = (D_{AM}^C - \alpha W_{AM}^C)^{-1}Y^{(0)}
\]
Algorithm 1 Attention Regularized Laplace Graph for Domain Adaptation (ARG-DA)

Input: Data $X$, Source domain labels $Y_S$, subspace dimension $k$, number of iterations $T$, regularization parameters $\lambda$ and $\alpha$

1. **Step 1:** Initialize the iteration counter $t=0$ and compute $M_0$ as in Eq (3).
2. **Step 2:** Initialize pseudo target labels $Y_T$ and projection space $A$.
3. **Step 3:** Initialize pseudo target labels $Y_T$ via a base classifier, e.g., 1-NN, based on source domain labels $Y_S$.

4. **Stage 1:** while Definition of Stage 1 do

   (1) Initialize $ARG^O$ based on the pseudo target labels and the common feature representation $X$ via Eq (11).

   (2) Solve the generalized eigendecomposition problem [15], [50] as in Eq (20) (replace $M^*$ by $M_0$) and obtain adaptation matrix $A$, then embed data via the transformation, $Z = A^T X$.

5. **Stage 2:** while Definition of Stage 2 do

   (1) Initialize $ARG^C$ based on the pseudo target labels and the common feature representation $X$ via Eq (21).

   (2) Solve the generalized eigendecomposition problem [15], [50] as in Eq (20) (replace $M^*$ by $M_0$) and obtain adaptation matrix $A$, then embed data via the transformation, $Z = A^T X$.

   (3) Construct the label matrix $Y_P$ as in Eq (16), then obtain $Y_{final}$ by solving Eq (21).

   (4) Update pseudo target labels $Y_{final}^{(t+1)} = [(n_s + 1 : (n_s + n_t))$;

6. **while not converged and $t < T$ do

   (5) Compute $M_{c}$ as in Eq (5).

   (6) Compute $M^*$ as defined in Eq (17) via updated $Y_T$.

7. **Step 5:** Stage 1.

8. **Step 6:** Stage 2.

9. **Step 7:** $t = t + 1$; Return to Step 3;

Output: Adaptation matrix $A$, embedding $Z$. Target domain labels $Y_T^{(t)}$

To sum up, at a given iteration, SP1 and SP2 can be dynamically optimized to search for the optimal solution and ensure the Attention Regularized Laplace Graph-based DA.

The complete learning algorithm is summarized in Algorithm 1 - ARG-DA.

D. Kernelization

The proposed ARG-DA method is extended to nonlinear problems in a Reproducing Kernel Hilbert Space via the kernel mapping $\phi: x \rightarrow \phi(x)$, or $\phi(X) : \{\phi(x_1), \ldots, \phi(x_n)\}$, and the kernel matrix $K = \phi(X)^T \phi(X) \in \mathbb{R}^{n \times n}$. We utilize the representer theorem to formulate the Kernel ARG-DA as:

$$
\min_{A^T, \lambda} \left\{ \sum_{c=1}^{C} tr(A^T K M^* + ARG^O K^T A) + \lambda \|A\|^2_T \right\} \\
+ \mu \left( \sum_{j=1}^{T} \sum_{i=1}^{C} \|Y_{ij}^{(t)} - Y_{ij}^{(0)}\| + \|Y_T^{(ARG^C)} Y\| \right)
$$

s.t. $M^* = M_0 + \sum_{c=1}^{C} (M_c) - (M_{S\rightarrow T} + M_{T\rightarrow S} + M_{S\rightarrow S})$, $A^T K H K^T A = I$, $\text{ARG}^O_{WA} = \text{W}^O \cdot \text{AM}$, $\text{ARG}^C_{WM} = \text{W}^C \cdot \text{AM}$

IV. EXPERIMENTS

A. Benchmarks and Features

In this paper, as shown in Fig.8, we evaluate the proposed ARG-DA using standard benchmarks, i.e., USPS [28]+MINIST [35], COIL20 [48], PIE [48], office+Caltech [48], and SVHN-MNIST [4], and compare it with state-of-the-art DA methods. Following the experimental settings as reported in the previous works [4], [10], [17], [43], [54], we construct 37 datasets for different image classification tasks. Due to space limitations, detailed explanations of the proposed datasets can be found in our previous research [52].

B. Baseline Methods

The proposed ARG-DA method is compared with thirty-three methods from the literature, which can be differentiated according to whether they are built on the deep learning-based framework:

- **Shallow methods:** (1) 1-Nearest Neighbor classifier (NN); (2) Principal Component Analysis (PCA); (3) GFK [18]; (4) TCA [58]; (5) TSL [71]; (6) JDA [48]; (7) ELM [77]; (8) AELM [77]; (9) SA [14]; (10) mSDA [5]; (11) TJM [49]; (12) RTML [10]; (13) SCA [17]; (14) CDML [80]; (15) LTL [70]; (16) LSR [85]; (17) KPCA [67]; (18) JGSA [88]; (19) CORAL [72]; (20) RVDLR [29]; (21) LPJT [37]; (22) DGA-DA [53]; (23) GEF through its different variants, GEF-PCA, GEF-LDA, GEF-LMNN, and GEF-MFA [6].

- **Deep methods:** (24) AlexNet [34]; (25) DAH [79]; (26) DANN [16]; (27) ADDA [75]; (28) LTRU [68]; (29) ATU [66]; (30) BSWD [65]; (31) DSN [4]; (32) DDC [76]; (33) DAN [46].

A direct comparison of the proposed ARG-DA using shallow features against DL-based DA methods could be unfair. To address this issue, in this paper, following the previous experimental settings as reported in DGA-DA, JGSA and BSWD, we make use of a deep feature, namely DeCAF6, as the input feature for a fair comparison with the DL-based DA methods. Whenever possible, we compared the performance scores of the thirty-three methods from the literature that are directly collected from their original papers or previous research [6], [17], [37], [53], [65], [75], [77], [88]. They are assumed to be their best performance.

C. Experimental Setup

Given the fact that the target domain has no labeled data under the experimental setting of UDA, it is not possible to tune a set of optimal hyper-parameters. In this paper, following the setting of previous research [48], [53], [85], we also evaluate the proposed ARG-DA by empirically searching the parameter space for the optimal settings. Specifically, the proposed ARG-DA method has three hyper-parameters, i.e., the subspace dimension $k$, and regularization parameters $\lambda$ and $\alpha$. In our experiments, we set $k = 200$ and $1) \lambda = 0.1$, and $\alpha = 0.9$ for USPS, MNIST, COIL20, and PIE, and for Office+Caltech, and SVHN-MNIST.
In our experiment, accuracy on the test dataset as defined by Eq. (24) is used as the performance measurement. It is widely applied in a series of previous research work, e.g., [46], [48], [54], [85], etc.

\[
\text{Accuracy} = \frac{\sum_{x \in D_T} \mathbb{1}[\hat{y}(x) = y(x)]}{|\sum_{x \in D_T}|} \tag{24}
\]

where \(D_T\) is the target domain treated as test data, \(\hat{y}(x)\) is the predicted label, and \(y(x)\) is the ground truth label for the test data \(x\).

As discussed in Sect.III-B, the core model of the proposed ARG-DA method is originated from DGA-DA but combines two optimization strategies, namely the Attention Regularized Laplace Graph (ARG) as defined in Eq.(11), and the Joint Manifold Alignment Strategy ('FEEL') as introduced in Sect.III-B.3. To gain insight into the proposed ARG-DA and the rationale w.r.t the ARG term and the FEEL strategy, we derive from Eq.(17) one intermediate partial model, namely DGA+A, to highlight the contribution of each optimization strategy to the final proposed ARG-DA:

- **DGA+A**: in this setting, we remove \(\text{ARG}^0\) from Eq.(17) and derive the DGA+A model, which only makes use of the ARG mechanism as defined in Eq.(11) but without the 'FEEL' strategy as defined in Sect.III-B.3. This setting makes it possible to understand the contribution of the proposed ARG mechanism when the different sub-domain adaptation is properly regularized to remedy one of the raised issues, namely 'Lack of Attention', by comparing it with its baseline DA model DGA-DA. Specifically, the mathematical formulations of both DGA+A and DGA-DA are depicted in Fig.15 for better clarity.

- **ARG-DA**: now our proposed ARG-DA improves DGA+A by further implementing the 'FEEL' strategy to additionally align the original feature manifold with DGA as baseline, it is now possible to highlight the contribution of the 'FEEL' strategy when all the manifold structures are further aligned.

**D. Experimental Results and Discussion**

1) Experiments on the CMU PIE Data Set: The CMU PIE database is a large face dataset containing 68 people with different pose, illumination, and expression variations. Fig.9 synthesizes the experimental results for DA using this dataset, where the best results are highlighted in red. The following observations can be drawn from Fig.9:

- by implementing the Attention Regularized Laplace Graph Enforced Adaptation model, DGA+A enhances discriminativeness across different sub-domains, thereby improving its baseline DGA-DA by 16 points, and achieves 81.25% accuracy. By making use of the FEEL strategy, ARG-DA additionally improves DGA+A by roughly 1 point and highlights the interest of unifying the alignment of different feature/label manifolds.

- Interestingly, the second best performing methods are GEF-LDA and GEF-PCA, respectively. They are both variants of GEF [6]. GEF-PCA, formalized as Eq.(25), improves the baseline JDA by 15 points by incorporating a specifically designed revision graph \(\Delta G\) in order to increase within-class compactness.

2) Experiments on the USPS+MNIST Dataset: Due to space limitation, the experimental results can be found in the supplemental materials.
results. As shown in Fig.10, thanks to the effectiveness of the previous research [53], Fig.10 reports the final experimental benchmark. Following the same experimental setting of DA domains, respectively, thereby generating a large-scale of 50k and 20k real-world images from two very different domains, thereby generating a large-scale of 50k and 20k real-world images from two very different domains, respectively. Furthermore, by additionally embracing SVHN-MNIST, the supplemental materials to space limitation, the experimental results can be found in the FEELing techniques, ARG-DA demonstrating its effectiveness.

Different from the other datasets, the COIL and PIE datasets. (accuracy w.r.t #iterations).

3) Experiments on the Office+Caltech-256 Datasets: Due to space limitation, the experimental results can be found in the supplemental materials.

4) Large-Scale Experiments on the SVHN-MNIST Dataset: Different from the other datasets, SVHN-MNIST is composed of 50k and 20k real-world images from two very different domains, respectively, thereby generating a large-scale DA benchmark. Following the same experimental setting of previous research [53], Fig.10 reports the final experimental results. As shown in Fig.10, thanks to the effectiveness of the ARG term enforced manifold embeddings, DGA+A improves DGA-DA by 1 point. Furthermore, by additionally embracing the FEEL strategy ensured comprehensive manifold learning techniques, ARG-DA further improves DGA+AA by an accuracy of 1 point. In Fig.10, ARG-DA achieves the best performance, even in comparison with a series of state-of-the-art deep learning-based DA methods, thereby further demonstrating its effectiveness.

5) Experiments on the COIL 20 Dataset: The COIL dataset (see fig.8) features the challenge of pose variations between the source and target domains. Even though DGA-DA already achieves 100.00% accuracy on the COIL dataset, we still benchmark our proposed ARG-DA in order to see how the proposed ARG term and FEEL strategy impact the best baseline performance. Fortunately, both DGA+A and ARG-DA achieve 100.00% accuracy and verify the robustness of the proposed ARG term and FEEL strategy.

E. Empirical Analysis

Although the proposed ARG-DA achieves state-of-the-art performance over 37 DA tasks through 7 datasets, an interesting question is how fast the proposed method converges (sect.IV-E.2), as well as its sensitivity w.r.t. its hyperparameters (Sect.IV-E.1). Additionally, in Sect.IV-E.3, we also aim to explore the robustness of the proposed FEEL strategy (Sect.III-B.3) without the collaboration of the ARG term.

1) Sensitivity of the proposed ARG-DA w.r.t. hyper-parameters: Three hyper-parameters, namely $k$, $\alpha$, and $\lambda$, are used in the proposed DA method.

- $k$ denotes the dimension of the searched shared latent feature subspace, which determines the structure of low-dimension embedding. Obviously, the larger $k$, the more the shared subspace can afford complex distributions, but at the cost of increased computation complexity.

- $\lambda$ defined in Eq.(23) aims to regularize the projection matrix $A$ to avoid over-fitting the chosen shared feature subspace w.r.t. both the source and the target domain. Increasing $\lambda$ reduces the risk of model over-fitting but shrinks the diversity of the functional learning representations.

- $\alpha = \frac{1}{\Gamma+\mu}$ as defined in Eq.(23) is a trade-off parameter, which balances $\text{LSmC}$ term (Eq.(15)) and label space embedding (Eq.(14)). Thus, increasing $\alpha$ improves the effectiveness of label space embedding, while reducing the contribution of label smoothness regularization.

Sensitivity of $k$: In Fig.13, we plot the classification accuracies of the proposed DA method w.r.t. different values of $k$ on the PIE dataset. As shown in Fig.13, the subspace dimensionality $k$ varies with $k \in \{20, 40, 60, 80, 100, 150, 200, 300, 400\}$. The proposed two DA variants, namely, DGA+A and ARG-DA, remain stable w.r.t. a wide range of $k \in \{60 \leq k \leq 400\}$. Finally, in this paper, we set $k = 200$ to balance efficiency and accuracy.

Sensitivity of $\lambda$ and $\alpha$: we also study the sensitivity of the proposed DGA+A and ARG-DA methods with a wide range of hyper-parameter values, i.e., $\alpha = (0.01, 0.05, 0.1, 0.5, 0.9, 0.95, 0.99)$ and $\lambda = (0.001, 0.002, 0.005, 0.1, 1, 5, 10)$. We plot in Fig.14 the results on COIL1 $\rightarrow$ COIL2 and MNIST $\rightarrow$ USPS datasets on both methods with $k$ held fixed at 200. As shown in Fig.14, the proposed DGA+A and ARG-DA display their stability as the resultant classification accuracy remains roughly the same despite a wide range of $\lambda$ and $\alpha$ values, thereby suggesting that
the proposed methods can easily search the proper candidate for these hyper-parameters.

2) Convergence Analysis: Another interesting question is to see when the proposed methods achieve their best performance w.r.t. the number of iterations $T$. For this purpose, in Fig.11, we carry out a convergence analysis of the proposed DGA+A and ARG-DA models, using the pixel value features on both the PIE and COIL datasets. Subsequently, Fig.11 reports 6 cross DA experiments ($COIL1 \rightarrow 2$, $COIL2 \rightarrow COIL1 \ldots PIE-27 \rightarrow PIE-5$) with the number of iterations $T = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$. As can be seen in Fig.11, both DGA+A and ARG-DA converge within 3−5 iterations when performing optimization over the two datasets. On the other hand, ARG-DA requires slightly more computational time per iteration than DGA+A, due to the computation of the additionally designed ARG$^G$, which discriminatively aligns different feature spaces. However, thanks to the FEEL strategy, the convergence curve displayed by ARG-DA is flatter in comparison with DGA+A.

3) Robustness of the FEEL-based comprehensive manifold learning: The FEEL strategy proposed in the ARG-DA model consists in aligning the manifold structures (MS) across the following groups of feature/label spaces:

- (G1): manifold structure across the original feature space ($X$) and the optimized common feature space ($A \ast X$);
- (G2): manifold structure across the optimized feature space ($A \ast X$) and the label space ($Y$);

Alignment of G1 or G2 in DA has already shown their effectiveness in a series of current research work, thereby encouraging hybridization of the previous two approaches by aligning jointly G1 and G2 within a unified optimization framework, as is the case in the proposed FEEL strategy. Even though the contribution of the FEEL strategy has already been highlighted by contrasting ARG-DA with its baseline model, DGA+A, over various datasets, we are still curious about the effectiveness of the FEEL strategy when it is disentangled from ARG regularization as in our previously proposed DGA-DA model. For this purpose, we derive a novel partial model, namely, DGA+F, which is formulated in Eq.(26)

$$\min \left( \sum_{i=1}^{C} tr(A^TM + L')K^2A) + \lambda \|A\|^2 + \mu \left( \sum_{i=1}^{C} \sum_{j=1}^{N} \|Y'_{ij} - Y''_{ij}\| + Y'LY \right) \right)$$

where $L'$ and $L$ intend to align the manifold structures of G1 and G2, respectively. To gain insight into the proposed FEEL strategy, we draw the general optimization framework in Fig.15, where four different experimental settings, e.g.,

- DGA-DA, DGA+F, DGA+A, and ARG-DA, are clearly distinguished by varying Term.1 and Term.2 of the optimization framework.

Fig.16 shows the performance of the four experimental settings using the pixel value features on both the USPS+MNIST and COIL datasets. As shown in Fig.16, the following can be observed:

- Mathematically, DGA+F is different from DGA-DA by updating Term.1 as $L'$, which intends to additionally align the manifold structure across G1. However, DGA+F underperforms in comparison with its baseline, DGA-DA, both on the USPS+MNIST and COIL datasets.
- Under the condition of using the ARG term-based regularization, ARG-DA improves DGA+A by incorporating the FEEL strategy, which can be seen on both the USPS+MNIST and COIL datasets.

To conclude, the FEEL strategy cannot be simply incorporated into the DGA-DA model. Due to the manifold structure divergence caused by domain shift, brutally aligning the manifold structure across G1 hurts the final DA performance. In this paper, by collaborating with the ARG term, the FEEL strategy is seamlessly embedded into the DGA+A model to dynamically align the manifold structure across different feature/label spaces and eventually achieve comprehensive manifold learning.

V. CONCLUSION

In this paper, we proposed a novel unsupervised DA method, namely Attention Regularized Laplace Graph-based Domain Adaptation (ARG-DA), which provides new perspectives, i.e., class aware attention and manifold alignment, to further explore Laplace graph-based DA methods. By weighting the importance of different sub-domain adaptations, we designed the Attention Regularized Laplace Graph to enforce the attention aware DA model DGA+A. Moreover, by unifying knowledge across different feature/label spaces, we improved DGA+A by incorporating the FEEL strategy for comprehensive manifold learning, thereby providing our final ARG-DA. Comprehensive experiments using the standard benchmark in DA showed the effectiveness of the proposed method, which consistently outperforms state-of-the-art DA methods. Future work includes embedding of the proposed ARG-DA into the paradigm of deep learning in order to improve the real computer vision applications, e.g., detection, segmentation, and tracking, etc.

REFERENCES

[1] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.
A. M. Martínez and A. C. Kak, “PCA versus LDA,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 7, pp. 3403–3417, Jul. 2015.

Y. Lu, L. Luo, D. Huang, Y. Wang, and L. Chen, “Knowledge transfer in vision recognition: A survey,” ACM Comput. Surv., vol. 53, no. 2, p. 37, 2020.

L. Luo, L. Chen, S. Hu, Y. Lu, and X. Wang, “Discriminative and geometry-aware unsupervised domain adaptation,” IEEE Trans. Cybern., vol. 50, no. 9, pp. 3914–3927, Sep. 2020.

L. Luo, X. Wang, S. Hu, and L. Chen, “Robust data geometric structure aligned close yet discriminative domain adaptation,” 2017, arXiv:1705.08620.

L. Luo, X. Wang, S. Hu, C. Wang, Y. Tang, and L. Chen, “Close yet distinctive domain adaptation,” 2017, arXiv:1704.04235.

A. M. Martínez and A. C. Kak, “PCA versus LDA,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 2, pp. 228–233, Feb. 2001.

Y. A. Ng, I. M. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” in Advances in Neural Information Processing Systems, vol. 14. T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA, USA: MIT Press, 2002, pp. 849–856.

S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” IEEE Trans. Neural Netw., vol. 22, no. 2, pp. 199–210, Feb. 2010.

S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Oct. 2009.

P. P. Busto and J. Gall, “Open set domain adaptation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 754–763.

V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, “Visual domain adaptation: A survey of recent advances,” IEEE Signal Process. Mag., vol. 32, no. 3, pp. 53–69, May 2015.

Z. Pei, Z. Cao, M. Long, and J. Wang, “Multi-adversarial domain adaptation,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018.

M. Pilanci and E. Ural, “Domain adaptation on graphs by learning aligned graph bases,” IEEE Trans. Knowl. Data Eng., vol. 34, no. 2, pp. 587–600, Feb. 2022.

S. Qian, T. Zhang, R. Hong, and C. Xu, “Cross-domain collaborative learning in social multimedia,” in Proc. 23rd ACM Int. Conf. Multimedia, Oct. 2015, pp. 99–108.

A. Rozantsev, M. Salzmann, and P. Fua, “Beyond sharing weights for deep domain adaptation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 4, pp. 801–814, Apr. 2019.

K. Saito, Y. Ushiku, and T. Harada, “Asymmetric tri-training for unsupervised domain adaptation,” in Proc. 34th Int. Conf. Mach. Learn. (ICML), Sydney, NSW, Australia, Aug. 2017, pp. 2988–2997.

B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” Neural Comput., vol. 10, no. 5, pp. 1299–1319, Jul. 1998.

O. Sener, H. O. Song, A. Saxena, and S. Savarese, “Learning transferrable representations for unsupervised domain adaptation,” in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 2110–2118.

L. Shao, F. Zhu, and X. Li, “Transfer learning for visual categorization: A survey,” IEEE Trans. Neural Netw. Learn. Syst., vol. 26, no. 5, pp. 1019–1034, May 2014.

M. Shao, D. Kit, and Y. Fu, “Generalized transfer subspace learning through low-rank constraint,” Int. J. Comput. Vis., vol. 109, nos. 1–2, pp. 74–93, 2014.

S. Si, D. Tao, and B. Geng, “Bregman divergence-based regularization for transfer subspace learning,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 7, pp. 929–942, Jul. 2010.

B. Sun, J. Feng, and K. Saenko, “Return of frustratingly easy domain adaptation,” in Proc. AAAI, vol. 6, 2016, p. 8.