Simultaneous Semantic Alignment Network for Heterogeneous Domain Adaptation

Shuang Li[1], Binhui Xie[1], Jiashu Wu[2], Ying Zhao[1], Chi Harold Liu[1,*], Zhengming Ding[3]

[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China.
[2]School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia.
[3]Department of Computer, Information and Technology, Indiana University-Purdue University Indianapolis.
{shuangli,binhuixie,yingzhao,chiliu}@bit.edu.cn,jiashu@student.unimelb.edu.au,zd2@iu.edu

ABSTRACT

Heterogeneous domain adaptation (HDA) transfers knowledge across source and target domains that present heterogeneities e.g., distinct domain distributions and difference in feature type or dimension. Most previous HDA methods tackle this problem through learning a domain-invariant feature subspace to reduce the discrepancy between domains. However, the intrinsic semantic properties contained in data are under-explored in such alignment strategy, which is also indispensable to achieve promising adaptability. In this paper, we propose a Simultaneous Semantic Alignment Network (SSAN) to simultaneously exploit correlations among categories and align the centroids for each category across domains. In particular, we propose an implicit semantic correlation loss to transfer the correlation knowledge of source categorical prediction distributions to target domain. Meanwhile, by leveraging target pseudo-labels, a robust triplet-centroid alignment mechanism is explicitly applied to align feature representations for each category. Notably, a pseudo-label refinement procedure with geometric similarity involved is introduced to enhance the target pseudo-label assignment accuracy. Comprehensive experiments on various HDA tasks across text-to-image, image-to-image and text-to-text successfully validate the superiority of our SSAN against state-of-the-art HDA methods. The code is publicly available at https://github.com/BIT-DA/SSAN.

CCS CONCEPTS

• **Computing methodologies** → **Transfer learning**; *Neural networks*; *Semi-supervised learning settings*.

KEYWORDS

Heterogeneous Domain Adaptation; Multimodal Alignment; Cross-domain Subspace Learning; Neural Network

*Corresponding author.

Figure 1: Examples of heterogeneous domain adaptation from left to right are 1) text-to-image classification with distinct data modalities; 2) object recognition with different feature dimensions; 3) cross-lingual document categorization.

1 INTRODUCTION

A vast majority of learning scenarios require sufficient amounts of labeled training data to achieve promising classification and generalization performance in a target domain. However, due to the expensive cost of data collection and laborious annotation for each problem of interest, it is crucial to transfer the enriched knowledge from an auxiliary domain (a.k.a., source domain) to facilitate learning a robust model for a target domain [37]. Towards this goal, a plenty of domain adaptation (DA) techniques have been successfully applied in various multimedia tasks such as multimodal learning [35, 40, 43], visual object recognition [13, 28, 44], and text categorization [4, 9, 16]. It is worth noting that most existing DA efforts are based on the assumption that the data from different domains are represented by the same type of features [3, 14, 26, 31, 33, 49, 60]. In other words, they cannot cope with the situation where the source and target domain data do not share the same feature representations, which is frequently encountered in real-world applications. Figure 1 presents three examples that involve data from different feature spaces. Take cross-lingual document categorization as an example, documents in English will not share the same feature space with those in Spanish, owing to distinct vocabularies varied among languages.

Heterogeneous domain adaptation (HDA) attempts to address the task of associating cross-domain data observed in separate feature spaces, which vary greatly from one dataset or domain to

another [12, 21]. In the literature, most current approaches focus primarily on learning one or two feature transformations to eliminate the heterogeneities in the data and hence facilitate the divergence minimization between domains. Formally, one can map data from one domain to the other [22, 24, 47, 53, 58] or map data into a common feature subspace [30, 42, 43, 52, 54]. For instance, Kulis et al. [24] first proposed an asymmetric regularized cross-domain transformation (ARC-t) to learn asymmetric and non-linear transformations with label information guarantees. While Shi et al. [42] presented a heterogeneous spectral mapping (HeMap), which derives two projection matrices based on spectral embedding.

Despite the great success achieved by these HDA methods, they would suffer from a major limitation: the instances of different categories are confounded together during the brute-force alignment process. As a result, they fail to produce discriminative features so that how to appropriately leverage the semantic properties underlying the data remains a significant issue. On one hand, the semantic correlations between classes have not been uncovered. The correlation relationship such as *laptop* should be more semantically similar to *monitor* than *mug* is useful to guide better alignment between domains and hence worth excavating. On the other hand, even if semantic alignment for traditional domain adaptation has received increasing attention [48, 55], to our best knowledge, semantic alignment has not been applied to address HDA problems due to separate feature spaces between domains, which becomes even harder to tackle when the labeled target data are limited.

To overcome the aforementioned challenges, we take a different tact and present a *Simultaneous Semantic Alignment Network (SSAN)* for HDA problems. In this work, we pursue how to contemporaneously transfer the semantic correlation knowledge of the original data as much as possible and yield discriminative and domain-invariant feature transformations. More precisely, to deal with the data that present heterogeneities, we first build two non-linear feature encoders using neural networks (one for each domain), which allow us to generate a common feature subspace. This common feature subspace can be used to carry out diverse features with arbitrary dimensionality. Hereafter, we utilize all labeled source data to train a shared classifier with the standard supervised classification loss. Meanwhile, with the observation that the classifier shares the similar prediction distribution for the same class between the labeled source and target data, we propose an *implicit semantic correlation* loss, which extends knowledge distillation [5, 20] to the HDA scenario. Thus, we can achieve alignment of the semantic correlations among categories across domains.

However, when optimizing the above two objectives to derive discriminative structures of supervised data in both domains, the unlabeled target data are failed to be fully excavated. As noted in [30], exploring unlabeled target instances during training would be helpful for HDA problems. To this end, we first strive to assign pseudo-labels for unlabeled target instances by ensemble predictions that reach consensus by both neural networks (the shared classifier) and the geometric similarity mechanism. In light of this, SSAN will automatically strengthen the confidence of pseudo-labels by using this pseudo-label refinement process. Furthermore, to reduce the bias caused by false pseudo-labels, we design a triplet-centroid alignment mechanism to match the centroids of transformed source, target and source-target combined features in each category, named

as *explicit semantic alignment*. Ultimately, for the purpose of more stable adaptation, we also consider a simple yet efficient way to facilitate distribution alignment with a domain confusion loss.

In summary, the contributions of our work are highlighted below:

- To our best knowledge, this is the first time to successfully utilize the semantic properties within data to solve HDA problems, promoting better adaptation capability.
- We present a novel Simultaneous Semantic Alignment Network, named SSAN, to consider and excavate the discriminative semantic structures of distinct distributions in HDA.
- Comprehensive evaluations on transfer tasks of text-to-image, image-to-image and text-to-text demonstrate the competence of our model, exceeding state-of-the-art HDA approaches.

2 RELATED WORK

Leveraging the knowledge extracted from the source domain, traditional domain adaptation [6, 27, 29, 32, 36, 61] aims to tackle the task of associating the homogeneous data described by the same type of features across distinct domains. Another branch of works follows the heterogeneous setting, where the data are not only drawn from dissimilar distributions but also represented by features with different dimensions or modalities, making it challenging to leverage knowledge obtained from source data to assist the target learning task. Here, we delve into heterogeneous domain adaptation works. In recent years, heterogeneous domain adaptation (HDA) has attracted increasing attention [22, 25, 30, 57]. The proposed SSAN distinguishes from existing HDA methods in two aspects:

Implicit Semantic Correlation Transfer. When bridging the gaps across domains, existing methods only leverage information extracted from the feature-level. Wang and Mahadevan [51] proposed a manifold alignment method (DAMA) to preserve label information during the alignment process. Duan et al. [12] presented Heterogeneous Feature Alignment (HFA), which incorporates augmented features and the standard SVM for performing recognition. Li et al. [30] extended HFA to a semi-supervised version (SHFA) by utilizing unlabeled target data during training. To learn a domain-invariant representation via an asymmetric category-independent transform, Hoffman et al. [21] proposed Max-Margin Domain Transforms (MMDT). Zhou et al. [59] presented Sparse Heterogeneous Feature Representation (SHFR) to cast the learning of feature mappings as a compression sensing problem. Yan et al. [56] put forward a semi-supervised entropic Gromov-Wasserstein discrepancy (SGW) approach to learn optimal transport from source to target domain features. Inspired by recent advances in deep learning, Chen et al. [7] proposed Transfer Neural Trees (TNT) with stochastic pruning to solve feature mapping and facilitate adaptation.

However, none of these methods utilize the semantic correlations contained in the predictions which can effectively guide the alignment process and facilitate better transferability.

Explicit Semantic Alignment. There exist several methods that apply pseudo-labels of unlabeled target instances to enforce the semantic alignment across domains. To learn representative cross-domain landmarks to derive proper feature subspace that eliminates the domain divergence, Tsai et al. [46] presented Cross-Domain Landmark Selection (CDLS). Similarly, Hsieh et al. [23] considered Generalized Joint Distribution Adaptation (G-JDA), which jointly

E_S: Source Encoder E_T: Target Encoder   C: Shared Classifier   D: Domain Discriminator   (LA) Label Assignment
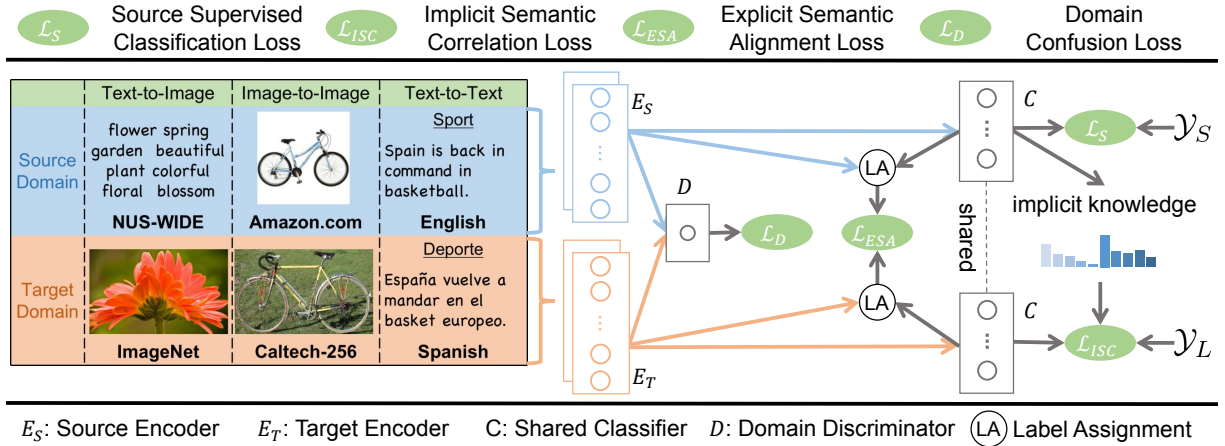
Figure 2: Approach overview. We employ an implicit semantic correlation loss over labeled target data to imitate the learned source prediction distributions. Simultaneously, we explicitly restrict the centroids of source, target and the combination of source and target in each class to be closer. To take advantage of the unlabeled target instances in the alignment process, we assign pseudo-labels by the consistent predictions between neural network (the shared classifier) and geometric similarity mechanism. Besides, we apply a domain confusion loss to learn domain-invariant feature representations.

matches both marginal and conditional distributions for adaptation and classification. All of this, unfortunately, are vulnerable to the error accumulation as the pseudo-label accuracy is not guaranteed, which is an important origin of performance reduction. Recently, to avert false pseudo-labels introduced by traditional hard label assignment, Yao et al. [57] proposed a Soft Transfer Network (STN) to adopt a soft-label strategy during class-level alignment.

In contrast, we introduce a geometric similarity pseudo-label refinement mechanism to automatically strengthen the confidence of pseudo-labels. With the objective of learning a more robust model, we further design a triplet-centroid alignment mechanism to match the centroids of transformed source, target and source-target combined features for each category in a progressive manner.

3 THE PROPOSED ALGORITHM

3.1 Preliminary and Motivation

For heterogeneous domain adaptation (HDA) with a semi-supervised setting, we have one labeled source domain and one scarcely labeled target domain. To be concrete, let $\mathcal{D}_S = \{\mathcal{X}_S, \mathcal{Y}_S\} = \{(\boldsymbol{x}_{s_i}, y_{s_i})\}_{i=1}^{n_s}$ denote a set of training instances of the source domain, where $\boldsymbol{x}_{s_i} \in \mathbb{R}^{d_s}$ denotes the i-th example with d_s-dimensional features and $y_{s_i} \in \{1, \ldots, K\}$ represents the corresponding label. Similarly, let $\mathcal{D}_L = \{\mathcal{X}_L, \mathcal{Y}_L\} = \{(\boldsymbol{x}_{l_i}, y_{l_i})\}_{i=1}^{n_l}$ and $\mathcal{D}_U = \{\mathcal{X}_U\} = \{\boldsymbol{x}_{u_i}\}_{i=1}^{n_u}$ denote labeled and unlabeled instances of the target domain, respectively, where $\boldsymbol{x}_{l_i}, \boldsymbol{x}_{u_i} \in \mathbb{R}^{d_t}$ and $y_{l_i} \in \{1, \ldots, K\}$. Notably, $d_s \neq d_t$ and $n_l \ll n_u$. In a nutshell, our ultimate goal is to produce a shared classifier C that can be applied to encoded features $\boldsymbol{f}(\boldsymbol{x}_i)$ of source and target data and correctly predict labels of the unlabeled target instances. For simplicity, we define that $\boldsymbol{f}_S(\boldsymbol{x}_i) = E_S(\boldsymbol{x}_i)$ if $\boldsymbol{x}_i \in \mathcal{X}_S$ and $\boldsymbol{f}_T(\boldsymbol{x}_i) = E_T(\boldsymbol{x}_i)$ if $\boldsymbol{x}_i \in \mathcal{X}_L \cup \mathcal{X}_U$, where E_S and E_T are source and target feature encoders, respectively.

Generally, most prior researches in HDA are devoted to associating heterogeneous data by learning a domain-invariant feature subspace. However, a drawback of these methods is that they ignore the semantic information contained in instances. Even though some

methods enforce the alignment between domains in a brute-force manner, they may still distort the features with the same class label due to the strong heterogeneities present across domains, which will damage the discriminative structures. To alleviate this concern, we aim to build two feature encoders E_S, E_T for source and target inputs, respectively, to yield a domain-invariant and discriminative feature subspace. Besides the standard supervision loss, we also explore the implicit knowledge learned from the labeled source data, transferring the correlations among categories. To mitigate the negative transfer caused by falsely-assigned pseudo-labels, we design a geometric pseudo-label refinement mechanism which can leverage the geometric property of the original data to boost pseudo-label assignment accuracy. Furthermore, to learn semantic representations for unlabeled target data, we propose the explicit semantic alignment in a progressive way by matching centroids corresponding to each category in the source, target and source-target combined feature representations via refined target pseudo-labels.

3.2 Implicit Semantic Correlation Transfer

Deep learning has been introduced into HDA and achieves remarkable performance across multimodal transfer tasks [7, 43, 57]. In general, it is the premise to train a discriminative classifier by minimizing the empirical error on labeled source data. Technically, we represent the source supervised classification loss as follows:

$$\mathcal{L}_S = \mathcal{J}_{sup}(\mathcal{X}_S, \mathcal{Y}_S) = \frac{1}{n_s} \sum_{\boldsymbol{x}_i \in \mathcal{X}_S, y_i \in \mathcal{Y}_S} \mathcal{J}_{ce}\left(C(\boldsymbol{f}_S(\boldsymbol{x}_i)), y_i\right), \quad (1)$$

where $\mathcal{J}_{ce}(\cdot, \cdot)$ is the cross-entropy loss. By the same token, we could utilize the limited target labeled data to learn the parameters of the target network (target encoder E_T and the shared classifier C) according to Eq. (1). However, this often results in overfitting to the limited labeled target data, which suffers from performance degradation when predicting in the unlabeled target domain.

In spite of the fact that the source and target domains present huge heterogeneity, it is reasonable that the classifier should produce the similar categorical probability distribution for the same

category in both domains, which is referred to as implicit knowledge. This is intuitive, for instance, *laptop* should have a closer semantic correlation with *monitor* and *keyboard* than *bike* and *mug* no matter whether it comes from the source domain or the target domain. Thus, we realize that the source category implicit knowledge could promote the generalization performance of classifier on the target domain. Therefore, we regard the average over the probabilistic outputs of source instances in class k as the k-th teacher and denote it as $q^{(k)} \in \mathbb{R}^K$, which is the "soft label" for class k. Owing to plenty of labeled source data available, "soft labels" contains much valuable category correlations. To adequately capitalize on these correlations, we adopt a softmax with a higher temperature T to smooth the classifier activations, which could produce a softer probability distribution over classes. Thus, we define $q^{(k)}$ as

$$q^{(k)} = \frac{1}{|\mathcal{X}_S^{(k)}|} \sum_{x_i \in \mathcal{X}_S^{(k)}} softmax\left(\frac{C\left(f_S(x_i)\right)}{T}\right), \qquad (2)$$

where $\mathcal{X}_S^{(k)}$ denotes the set of source instances of class k and $|\cdot|$ represents the number of samples in the set. Given a labeled target instance, we could fine-tune the target network with "soft labels" to learn and transfer the semantic correlations from source to target domain. Hence, with the supervision of the learned "soft labels", the corresponding loss can be calculated as

$$\mathcal{J}_{soft}\left(\mathcal{X}_L, \mathcal{Y}_L\right) = -\frac{1}{n_l} \sum_{x_i \in \mathcal{X}_L, y_i \in \mathcal{Y}_L} q^{(y_i)^\top} \log p_i, \qquad (3)$$

where p_i is the probabilistic output for labeled target instance x_i, and $p_i = softmax\left(C(f_T(x_i))\right)$. The above loss can transfer the semantic correlations from the source network to the target network. To conclude, we further take the supervised loss of labeled target data into consideration and define the implicit semantic correlation loss as

$$\mathcal{L}_{ISC} = (1 - \alpha)\,\mathcal{J}_{sup}\left(\mathcal{X}_L, \mathcal{Y}_L\right) + \alpha\,\mathcal{J}_{soft}\left(\mathcal{X}_L, \mathcal{Y}_L\right). \qquad (4)$$

In this way, the target network can perform better generalization to these instances around class boundaries and capture the semantic correlations among categories in supervised data, which will achieve significant performance gains.

3.3 Explicit Semantic Alignment

Recall that we dedicate to learning discriminative representations for unlabeled target instances by aligning the conditional distributions of source and target domains. However, we do not have label information for unlabeled target instances. One feasible way is to utilize pseudo-labels [23, 46] predicted by the shared classifier directly, but there is no doubt that the adaptation capability would be impeded by falsely-assigned pseudo-labels, which will cause the accumulation of negative effect during the alignment process. To circumvent the uncertainty of pseudo-labels, we resort to uncovering the intrinsic geometric knowledge underlying data. Thus we design a geometric pseudo-label refinement mechanism to assist in assigning pseudo-labels for those instances that present geometric similarity to the category centroids of supervised data in the feature space. We first compute the centroid $\mu^{(k)} \in \mathbb{R}^{d_{common}}$ of category k in the labeled source and target domains, which is a mean vector of both encoded features in each category and is calculated by:

$$\mu^{(k)} = \frac{1}{|\mathcal{X}_S^{(k)} \cup \mathcal{X}_L^{(k)}|} \left(\sum_{x_i \in \mathcal{X}_S^{(k)}} f_S(x_i) + \sum_{x_j \in \mathcal{X}_L^{(k)}} f_T(x_j) \right), \qquad (5)$$

where $\mathcal{X}_L^{(k)}$ denotes the set of labeled target instances of class k. Then, a set of centroids $\{\mu^{(k)}\}_{k=1}^K$ are obtained and we use a geometric similarity metric to assign i-th unlabeled target instance with a geometric similarity label:

$$y_{u_i}^{<GS>} = \arg\max_k GS\left(f_T(x_{u_i}), \mu^{(k)}\right), \qquad (6)$$

where $GS(\cdot, \cdot)$ is the geometric relationship between two data points in the latent feature space and the Cosine similarity is chosen in this paper. Moreover, it is easy to get pseudo (predicted) label from the shared classifier C and we define as $y_{u_i}^{<NN>}$. With geometric similarity label and neural network label, an unlabeled target instance will only be selected and assigned with pseudo-label if $y_{u_i}^{<GS>} = y_{u_i}^{<NN>}$, which can boost pseudo-label assignment accuracy. Thus, we have $\mathcal{X}_T = \mathcal{X}_L \cup \hat{\mathcal{X}}_U$ and $\mathcal{Y}_T = \mathcal{Y}_L \cup \hat{\mathcal{Y}}_U$, where $\hat{\mathcal{X}}_U, \hat{\mathcal{Y}}_U$ are the sets of selected unlabeled target instances and their corresponding pseudo-labels.

Intuitively, few unlabeled target samples are likely to be correctly annotated in the early training phase. As the training evolves, more and more samples will potentially be assigned with an agreed pseudo-label. Ultimately, the majority of samples will be assigned with confident pseudo-labels. Note that this automatic mechanism does not require manual experience to set the threshold. This hard example mining process progressively encourages those instances with consistent predicted label to be accepted to participate in the explicit semantic alignment while filtering out instances without consistency. Accordingly, we are able to adequately learn semantic representations for unlabeled target instances using the designed explicit semantic alignment loss. Formally, there are three centroids of source, target, and combination of source and target in the common feature space for each class, constituting the triplet centroids:

$$\mu_S^{(k)} = \frac{1}{|\mathcal{X}_S^{(k)}|} \sum_{x_i \in \mathcal{X}_S^{(k)}} f_S(x_i), \quad \mu_T^{(k)} = \frac{1}{|\mathcal{X}_T^{(k)}|} \sum_{x_j \in \mathcal{X}_T^{(k)}} f_T(x_j),$$

$$\mu_{ST}^{(k)} = \frac{1}{|\mathcal{X}_S^{(k)} \cup \mathcal{X}_T^{(k)}|} \left(\sum_{x_i \in \mathcal{X}_S^{(k)}} f_S(x_i) + \sum_{x_j \in \mathcal{X}_T^{(k)}} f_T(x_j) \right), \qquad (7)$$

where $\mathcal{X}_T^{(k)}$ represents the subset of $\mathcal{X}_T$ including labeled target instances whose ground-truth labels are class k and unlabeled target instances with consentaneously assigned pseudo-labels of class k.

Once the centroids are calculated, they can facilitate semantic alignment and enhance better semantic consistency between domains. Drawing inspiration from the maximum mean discrepancy (MMD) technique which has been proved to be an effective measurement of the divergence between two domains [17, 38, 50]. Naturally, we employ the following explicit semantic alignment loss to learn more robust and discriminative representations between domains:

$$\mathcal{L}_{ESA} = \sum_{k=1}^K \left(\left\|\mu_S^{(k)} - \mu_T^{(k)}\right\|_2^2 + \left\|\mu_S^{(k)} - \mu_{ST}^{(k)}\right\|_2^2 + \left\|\mu_T^{(k)} - \mu_{ST}^{(k)}\right\|_2^2 \right). \qquad (8)$$

By minimizing this objective, the centroids of each category will be forced to be in close proximity in the encoded feature subspace, resulting in semantically consistent representations across domains.

Furthermore, the explicit semantic alignment loss enforces semantic consistency in local distributions, while ignoring the global distributions between domains. To this end, we additionally add a domain discriminator D with a single layer to distinguish whether the encoded features are from source or target domain while encoders are trained to fool D [15]. In other words, D simply performs

the binary classification task. Concretely,

$$\mathcal{L}_D = \frac{1}{n_s} \sum_{\boldsymbol{x}_i \in \mathcal{X}_S} \log \left(D(f_S(\boldsymbol{x}_i)) \right)$$

$$+ \frac{1}{n_l + n_u} \sum_{\boldsymbol{x}_j \in \mathcal{X}_L \cup \mathcal{X}_U} \log \left(1 - D(f_T(\boldsymbol{x}_j)) \right) . \quad (9)$$

The feature representations from encoders E_S, E_T are domain-invariant when this minimax game reaches an equilibrium.

3.4 Overall Formulation and Optimization

In summary, we exploit the implicit semantic knowledge in the prediction space for the preservation of category correlations and align features of both domains in each category to facilitate better alignment and transferability. To preserve semantic correlations and keep semantic consistency between domains, we design an implicit semantic correlation loss $\mathcal{L}_{ISC}$ to force the target network to learn the correlations of source predictions, and an explicit semantic alignment loss $\mathcal{L}_{ESA}$ with the help of geometric pseudo-label refinement mechanism to match the centroids of source, target and source-target combined features for each category. In addition, a domain confusion loss $\mathcal{L}_D$ is introduced to reduce the global domain divergence for improving stability. To this end, we present our overall objective function in a minimax scheme as:

$$\min_{C, E_S, E_T} \max_D \mathcal{L}_S + \mathcal{L}_{ISC} + \beta \mathcal{L}_{ESA} + \gamma \mathcal{L}_D, \quad (10)$$

where the hyper-parameters β and γ balance the influences of $\mathcal{L}_{ESA}$ and $\mathcal{L}_D$ on the optimization process, respectively.

4 EXPERIMENT

4.1 Datasets and Setup

NUS-WIDE [8] and **ImageNet** [10] are used for text-to-image transfer task. The former contains 269,648 images with tag information from Flickr, while the latter includes 5,247 synsets and 3.2 million images. We use the tag information of NUS-WIDE (N) and the image data of ImageNet (I) as the source domain (text) and the target domain (image), respectively. The lack of co-occurrence between text and image domain data makes target image classification a challenging cross-modal learning task. Based on the protocol in [7, 57], 8 overlapping categories of these two datasets are chosen. We pre-process the NUS-WIDE tag data by pre-training a 5-layer neural network with a softmax layer and extract the 4-th hidden layer as the 64-dimensional features for text data. We then follow [45] to extract 4096-dimensional DeCAF$_6$ features for image data. For the source domain, 100 texts per category in NUS-WIDE are chosen to form the labeled source data. As for the target domain (ImageNet), 3 images in each class are randomly sampled to be used as labeled target data and all remaining images are used for testing.

Office+Caltech-256 [18, 41]. The former is composed of 4,652 images belonging to 31 classes collected from three domains: Amazon (A), Webcam (W) and DSLR (D). The latter includes 30,607 images with 256 object classes from Caltech-256 (C). 10 overlapping classes are chosen to construct the Office+Caltech-256 dataset for image-to-image experiments [12, 46]. Furthermore, we consider three types of feature representations: 800-dimensional SURF features [2] and 4096-dimensional DeCAF$_6$ features [11], as well as 2,048-dimensional ResNet50 features [19]. Among these transfer tasks, all source domain images are utilized. As suggested by [7, 23,

46, 57], for the target domain, we randomly choose 3 images per class as the labeled target data, while all remaining images are set as unlabeled target data to be recognized. Note that DSLR is only viewed as the target domain due to the limited amount of images. To testify the effectiveness and robustness of our SSAN model against various HDA scenarios, two kinds of transfer tasks are conducted: (1) Transfer tasks across heterogeneous features within the same domain (i.e., A → A, C → C, W → W), where SURF → DeCAF$_6$ and DeCAF$_6$ → SURF are two settings adopted for source and target features, respectively. (2) Transfer tasks across heterogeneous features and distinct domains, in which SURF features are used as features for the source domain and DeCAF$_6$ or ResNet50 features are used as features for the target domain. Also, source and target data come from distinct domains. We evaluate our method on 9 adaptation scenarios: A → C, ..., W → C, W → D. By performing all these tasks, we can verify that our SSAN is effective and robust to tackle heterogeneities including diverse features and distinct domains.

Multilingual Reuters Collection [1]. This multilingual text categorization dataset consists of about 11,000 articles from 6 categories written in five different languages: English (EN), French (FR), German (GR), Italian (IT) and Spanish (SP). Following [12, 23, 30, 46], all the articles are represented by Bag-of-Words (BOW) with TF-IDF features, and then perform dimensionality reduction of features using PCA with 60% energy preserved. The reduced dimensions of five languages EN, FR, GE, IT and SP are 1,131, 1,230, 1,471, 1,041 and 807, respectively. We consider EN, FR, GR and IT as the source domain and SP as the target one, following [46, 57]. For the source domain, 100 articles per category are randomly selected to constitute the labeled source data. For the target domain, we randomly pick up m (i.e., m = 5, 10, 15, 20) and 500 articles per category as labeled and unlabeled target data, respectively. By choosing different amounts of labeled data, we can investigate how the number of labeled target data affects the transferability.

Implementation Details. We implement our SSAN using the PyTorch framework [39]. Both the source and target feature encoders are two-layer neural networks, which apply Leaky ReLU [34] as the activation function following [57]. The classifier and the domain discriminator are both neural networks with a single layer. As for parameter settings, we empirically set α = 0.1, β = 0.004, γ = 0.01, T = 5 and the dimension of the common subspace d_{common} is set to 256. Parameter sensitivity is conducted in Section 4.3 to testify the robust performance on a wide range of parameter values.

4.2 Evaluation

We evaluate the performance of the SSAN by first considering SVMt and NNt as baseline methods, which simply train a support vector machine and a two-layer neural network with only the labeled target data, respectively. Then, we compare our method against several state-of-the-art HDA methods (e.g., MMDT [21], CDLS [46], TNT [7], STN [57] etc.) to demonstrate the potency of our method. We apply the classification accuracy of unlabeled target data as the evaluation metric, which is widely used in existing literature [30, 46, 57]. Note that partial reported results are copied from their published papers if the experimental setup is the same.

Text-to-Image transfer task: Table 1 presents the results of transfer task from NUS-WIDE to ImageNet. When performing alignment and classification on this task, our method achieves the best

Table 1: Classification results (%) with standard deviations for text-to-image transfer scenario.

S → T	SVMt	NNt	MMDT [21]	SHFA [30]	SGW [56]	CDLS [46]	G-JDA [23]	TNT [7]	STN [57]	SSAN
N → I	66.85±0.96	67.68±0.80	53.21±0.69	64.06±0.61	68.01±0.80	70.96±0.83	75.76±0.65	77.71±0.57	78.46±0.58	**80.22±0.91**

Table 2: Classification results (%) with standard deviations for cross-feature object recognition.

S → T	SVMt	NNt	DAMA [51]	MMDT [21]	SHFR [59]	SHFA [30]	G-JDA [23]	CDLS [46]	STN [57]	SSAN
				SURF to DeCAF$_6$						
A → A	88.66±0.50	90.00±0.33	87.40±0.50	89.30±0.40	87.10±1.20	88.60±0.30	92.30±0.20	91.70±0.20	92.19±0.83	**92.45±0.52**
C → C	77.31±1.10	79.56±0.50	73.80±1.20	80.30±1.20	73.20±1.60	78.20±1.00	86.70±0.50	81.80±1.10	82.92±0.92	**87.01±0.71**
W → W	89.32±1.10	91.42±0.73	87.20±0.70	87.30±0.80	87.50±1.30	90.00±1.00	89.40±0.90	95.20±0.90	95.43±0.78	**96.66±0.62**
				DeCAF$_6$ to SURF						
A → A	43.03±0.90	42.82±1.77	38.10±1.10	40.50±1.30	44.50±1.10	42.90±1.00	50.30±0.70	46.40±1.00	47.62±1.49	**52.91±0.96**
C → C	30.15±1.50	31.33±2.89	18.90±1.30	30.60±1.70	33.40±1.00	29.40±1.50	33.70±0.80	31.80±1.20	30.83±2.14	**37.24±0.95**
W → W	55.28±1.00	60.87±1.58	47.40±2.10	59.10±1.20	54.30±0.90	62.20±0.70	63.80±0.90	63.10±1.10	64.71±1.62	**69.81±0.88**

Table 3: Classification results (%) with standard deviations for cross-domain object recognition (SURF to DeCAF$_6$).

S → T	SVMt	NNt	MMDT [21]	SHFA [30]	CDLS [46]	SGW [56]	G-JDA [23]	TNT [7]	STN [57]	SSAN
A → C	79.64±0.46	81.03±0.50	75.62±0.57	71.16±0.73	78.73±0.49	79.88±0.53	86.60±0.17	85.79±0.42	88.21±0.16	**88.36±0.73**
A → D	92.60±0.71	92.99±0.63	91.65±0.83	95.16±0.36	94.45±0.59	93.43±0.67	90.67±0.65	92.04±0.76	96.42±0.43	**97.00±0.77**
A → W	89.34±0.94	91.13±0.73	89.28±0.77	88.11±1.01	91.57±0.81	90.26±0.84	94.09±0.67	91.26±0.72	**96.68±0.43**	96.31±0.64
C → A	89.13±0.39	89.60±0.33	87.06±0.47	85.49±0.51	86.34±0.74	89.03±0.37	92.49±0.12	92.35±0.17	93.03±0.16	**93.16±0.51**
C → D	92.60±0.71	92.99±0.63	91.46±0.85	94.25±0.50	90.43±0.79	93.43±0.67	88.62±0.76	92.67±0.80	96.06±0.50	**97.44±0.61**
C → W	89.34±0.94	91.13±0.73	89.11±0.76	89.47±0.90	88.60±0.80	90.26±0.84	92.64±0.54	92.98±0.75	**96.38±0.38**	95.87±0.72
W → A	89.13±0.39	89.60±0.33	87.00±0.47	88.83±0.45	87.51±0.44	89.02±0.37	92.28±0.15	92.99±0.14	93.11±0.16	**93.54±0.44**
W → C	79.64±0.46	81.03±0.50	75.44±0.59	79.66±0.52	77.30±0.71	79.85±0.53	84.82±0.38	86.28±0.51	87.22±0.45	**88.18±0.61**
W → D	92.60±0.71	92.99±0.63	91.77±0.83	95.31±0.63	92.72±0.75	93.43±0.67	95.87±0.41	94.09±0.88	96.38±0.57	**97.64±0.51**
Avg.	87.68±0.63	88.69±0.55	86.49±0.68	87.49±0.62	87.52±0.68	88.73±0.61	90.90±0.43	91.17±0.44	93.72±0.36	**94.17±0.97**

Table 4: Classification results (%) with standard deviations for cross-domain object recognition (SURF to ResNet50).

S → T	SVMt	NNt	CDLS [46]	STN [57]	SSAN
A→C	19.99±1.78	20.22±1.96	21.32±2.01	21.64±2.36	**24.07±2.00**
A→D	84.37±3.54	84.25±3.78	85.03±1.97	85.20±2.80	**91.81±3.19**
A→W	78.60±3.36	81.51±3.01	80.38±3.02	83.77±3.19	**88.47±1.21**
C→A	34.26±3.13	34.81±3.51	34.69±1.83	34.16±3.11	**41.71±2.07**
C→D	84.37±3.54	84.25±3.78	86.61±2.76	86.59±3.57	**91.22±2.35**
C→W	78.60±3.36	81.51±3.01	78.11±2.08	84.21±2.56	**86.24±2.04**
W→A	34.26±3.13	34.81±3.51	35.66±1.19	35.18±2.79	**37.36±1.98**
W→C	19.99±1.78	20.22±1.96	20.31±1.01	20.71±2.66	**23.50±2.19**
W→D	84.37±3.54	84.25±3.78	82.68±1.57	83.74±2.96	**93.62±2.27**
Avg.	57.65±3.02	58.43±3.14	58.31±1.94	59.47±2.89	**64.22±2.14**

performance among all methods. Compared with the best supervised method NNt, our SSAN achieves the performance boost of more than **12%**. In addition, our method also outperforms STN, the best-performed baseline, by a large margin. Due to the presence of challenging heterogeneities between this text to image adaptation task, even though the soft-label strategy is utilized in STN, it is still not enough to avert the negative effects of false pseudo-labels. Thus, it would be desirable to consider both semantic correlations between classes and progressively conduct semantic alignment. The performance gains achieved by our SSAN over other state-of-the-art HDA methods could indicate the effectiveness of SSAN.

Image-to-Image transfer tasks: The classification results of transfer tasks across features within the same domain are summarized in Table 2. Based on the experimental results from Table 2, it is

desirable that our SSAN obtains significant improvements over the previous methods on all six tasks. The convincing results successfully verify that our model could effectively tackle features where heterogeneities present and is robust regarding domains. On the other hand, the classification results of transfer tasks across features and domains are shown in Table 3 and Table 4. In terms of transferring between shallow features (SURF) and diverse deep features (DECAF$_6$ and ResNet50), our SSAN reaches the highest average accuracies of **94.17%** and **64.22%** among all the compared methods, indicating the robustness and effectiveness of SSAN when coping with diverse feature spaces. Ultimately, the above results support that our SSAN could capture enriched semantic information to facilitate learning more transferable feature representations.

Text-to-Text transfer task: In Figure 3, with respect to text-to-text transfer task, we compare SSAN with previous approaches based on the different number of labeled target instances per class. Our model outperforms the comparisons among all the tasks. For the most challenging tasks with only 5 labeled target instances being available, the average classification accuracy over all these four tasks (i.e., EN→SP$_5$, FR→SP$_5$, GE→SP$_5$, IT→SP$_5$) is **72.53%**, which outperforms the best HDA method, i.e., STN, by **2.68%**. Intuitively, we could observe that by increasing the number of labeled target instances, better alignment is facilitated and the performance of all methods on all tasks is boosted. It is also reasonable to see that most HDA methods can yield either comparable or superior performance compared with supervised methods, i.e., NNt and SVMt, which demonstrates that HDA methods can facilitate divergence minimization and positive transfer in all text-to-text transfer tasks.
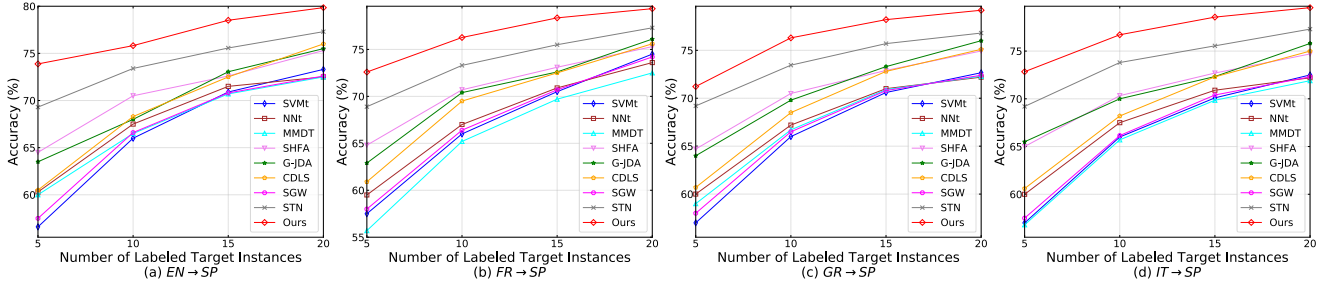
Figure 3: Classification results w.r.t. different number of labeled target instances per class on the multilingual Reuters dataset.

Table 5: Ablation study of SSAN.

Method/Task	$N \rightarrow I$	$C \rightarrow D$	$EN \rightarrow SP_5$
SSAN ($\alpha = 0$)	73.24±1.15	96.22±0.91	66.91±0.89
SSAN ($\beta = 0$)	69.21±1.80	96.09±0.78	64.39±1.96
SSAN ($\gamma = 0$)	77.59±1.65	96.30±0.86	71.98±1.01
SSAN (w/o T)	71.34±0.80	96.14±1.07	65.48±2.15
SSAN (w/o GS)	77.30±0.98	96.30±0.94	69.11±1.54
SSAN (full)	**80.22±0.91**	**97.44±0.61**	**73.68±0.50**

4.3 Insight Analyses

Ablation Study. Ablation study is conducted by evaluating several variants of SSAN: (1) SSAN ($\alpha = 0$), which removes $\mathcal{L}_{soft}$ from Eq. (4); (2) SSAN ($\beta = 0$), which removes the explicit semantic alignment; (3) SSAN ($\gamma = 0$), which turns off the $\mathcal{L}_D$; (4) SSAN (w/o T), which removes the temperature from our model (i.e., $T = 1$ in Eq. (2)); (5) SSAN (w/o GS), which assigns pseudo-labels for unlabeled target instances only by the neural network (shared classifier) predictions instead of the integration of predictions from both the neural network and the geometric similarity mechanism.

As shown in Table 5, the results are performed on randomly choosing $N \rightarrow I$, C (SURF) $\rightarrow$ D (DeCAF$_6$), and $EN \rightarrow SP_5$ as the representative tasks for text-to-image, image-to-image, and text-to-text transfer scenarios. We can clearly observe that the full SSAN outperforms all its variants by a large margin, which indicates that any one of these components plays an indispensable role and brings benefits to facilitate the positive transfer. Among all these components, implicit semantic correlation knowledge transfer and explicit semantic alignment bring the top two highest total gain of three tasks, which manifests the importance of transferring both explicit and implicit semantic knowledge. On the other hand, the performance gain contributed by the domain adversarial alignment in SSAN ($\gamma = 0$) is inferior to the semantic alignment. This observation suggests that despite the positive effect of domain alignment, it possesses less importance compared with semantic alignment components. With regard to SSAN (w/o T), it further indicates that higher temperature T would have a positive performance gain since setting $T = 1$ could not fully explore the correlations among categories during semantic transfer. Furthermore, the inclusion of geometric pseudo-label refinement enhances the performance by **2.88%** on average, which verifies that the geometric pseudo-label refinement can facilitate a more accurate pseudo-label assignment and boost the performance.

Significance Test (t-Test). To further validate the superiority of the SSAN, a statistical significance test is shown in Figure 4. A
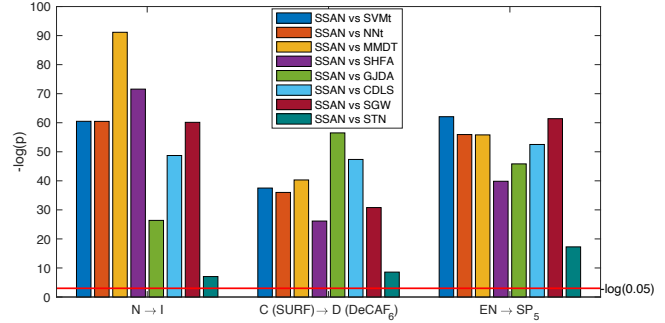

Figure 4: The significant test with 0.05 as the significance level is conducted to testify whether the performance gains of our SSAN compared with other baselines on three transfer tasks are statistically significant. The larger the value is, the more statistically significant the result indicates.

significant level of 0.05 is applied, and a p-value that is less than 0.05 indicates that the performance boost of our SSAN compared with another method is statistically significant. In Figure 4, we illustrate the $-\log(p)$ for each p-value, and a red line which indicates the base significance level of 0.05 ($-\log(0.05)$). The larger the value of $-\log(p)$ is, the more statistically significant the result indicates. Based on the results, it verifies the significant superiority of our SSAN compared with other comparable approaches in all three distinct transfer scenarios, which further demonstrates the effectiveness and robustness of the SSAN.

Parameter Sensitivity. To validate the parameter sensitivity of our model, we randomly take $N \rightarrow I$, C (SURF) $\rightarrow$ D (DeCAF$_6$), and $EN \rightarrow SP_5$ as the representative tasks. We vary parameters including α, β, γ, T and the dimension of common subspace d_{common} within their corresponding reasonable ranges, and plot the results in Figure 5. We can observe that the results demonstrate the robustness of SSAN under a wide range of parameter choices, which indicates that our model is stable and effective.

Implicit Semantic Correlation Visualization. We investigate the efficacy of our implicit semantic correlation knowledge and show it in Figure 6. The leftmost histograms illustrate the average prediction distribution of source instances that belong to the category of the shown images. They implicitly express the semantic correlation relationships among categories. By leveraging this implicit knowledge during the alignment process, the learner can learn the semantic correlations, for instance, the target monitor category can mimic a similar predictive distributions just as what the source monitor category has. Therefore, applying this implicit
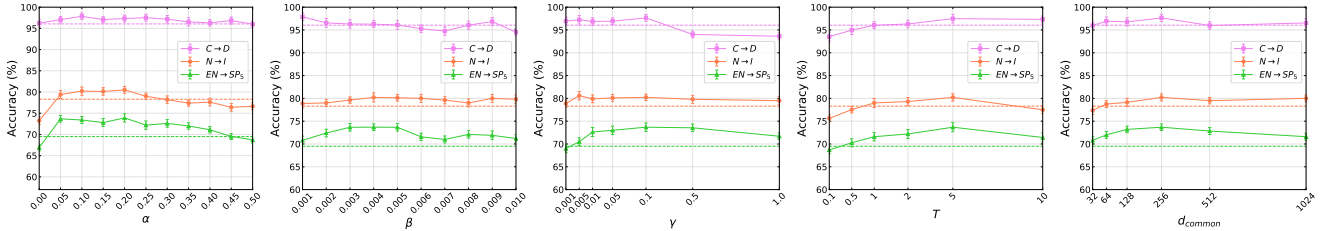
Figure 5: Parameter sensitivity on three random transfer tasks. The dashed lines show the corresponding best baseline results.
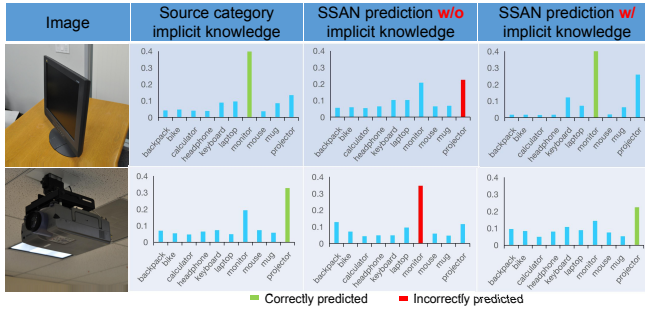

Figure 6: Two illustrative examples, their corresponding source semantic correlation knowledge, and predicted distributions without/with implicit knowledge are presented.

knowledge can transfer the correlation relationship implicitly contained in source predictions which will facilitate the alignment process and mitigate the possibility of category mismatches. Moreover, the implicit knowledge can correct some wrong predictions and hence prevent the model from being confused and learn a more robust classifier. The middle and the rightmost histograms show the predicted distributions of the shown image instances without using or using the implicit semantic correlation transfer, respectively. The results verify that by taking implicit knowledge into account, our SSAN can successfully classify both instances.

Geometric Similarity Pseudo-label Refinement. To examine the effectiveness of our geometric pseudo-label refinement mechanism, in Figure 7, pie charts are utilized to statistically summarize pseudo-label predictions of task C (SURF) → D (ResNet50) for three different phases during training, and some example images with their assigned pseudo-labels at epoch 500 are presented. The blue portion in pie charts represents relatively easy instances which are easy to be correctly predicted by both the neural network (NN) and the geometric similarity (GS) mechanism, which form our pseudo-label refinement procedure. The instances of this portion will be leveraged during the alignment process as the pseudo-labels assigned by two learners reaching a consensus. On the other hand, instances belonging to the grey and the yellow portion will be filtered out by our refinement procedure due to the disagreement of assigned pseudo-labels by two learners. Hence, hard instances like the projector and the instances that do not present strong geometric similarity like the headphone will not be assigned with a pseudo-label used for alignment. This effectively mitigates the negative effects caused by these instances during the alignment process. For those instances in the orange portion such as the laptop example, although both learners incorrectly predict their pseudo-labels, however, there is still a chance that the pseudo-labels will not be the same, which can still pare some hard instances down from
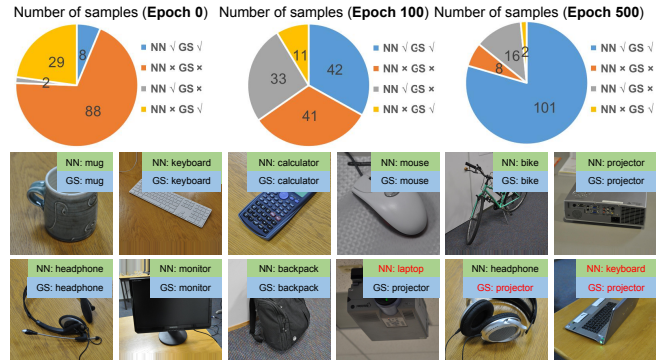

Figure 7: Top: Pie charts statistically summarize the amount of correctly and incorrectly assigned pseudo-labels predicted by NN (neural network) and GS (geometric similarity) during different phases of training on task C (SURF) → D (ResNet50). Tick and cross stand for correctly assigned pseudo-label and incorrectly assigned pseudo-label, respectively. Bottom: Example images with their assigned pseudo-labels at epoch 500 are presented. The green and light-blue tags represent NN and GS assigning pseudo-labels, respectively. Incorrect pseudo-label assignments are shown in red.

the alignment process. Unlike previous methods such as [23, 46] which directly utilize pseudo-labels assigned by NN without any refinement, our SSAN effectively refines pseudo-labels by considering geometric semantics. This ensemble procedure will raise the pseudo-label assignment accuracy and hence facilitate better fine-level semantic alignment.

5 CONCLUSION

In this paper, we propose a novel Simultaneous Semantic Alignment Network (SSAN) to address the HDA problems. Compared with existing HDA methods, our method simultaneously excavates both implicit and explicit semantic knowledge to facilitate alignment between cross-domain heterogeneous data. Implicit semantic knowledge facilitates achieving better preservation of categorical distributional correlations, while explicit triplet-centroid alignment procedure with the geometric semantic label refinement enforces cross-domain semantic alignment and consistency. The experimental results on a variety of cross-modal transfer tasks demonstrate the superiority of SSAN against several state-of-the-art HDA methods.

ACKNOWLEDGEMENTS

REFERENCES

[1] Massih R. Amini, Nicolas Usunier, and Cyril Goutte. 2009. Learning from Multiple Partially Observed Views -an Application to Multilingual Text Categorization. In *NeurIPS*. MIT Press, 28–36.
[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *ECCV*. Springer, 404–417.
[3] Shai Bendavid, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of Representations for Domain Adaptation. In *NeurIPS*. MIT Press, 137–144.
[4] John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain Adaptation with Structural Correspondence Learning. In *EMNLP*. ACL, 120–128.
[5] Cristian Buciluundefined, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model Compression. In *ACM SIGKDD*. ACM, 535–541.
[6] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. 2019. Progressive Feature Alignment for Unsupervised Domain Adaptation. In *CVPR*. IEEE, 627–636.
[7] Weiyu Chen, Tzuming Harry Hsu, Yaohung Hubert Tsai, Yuchiang Frank Wang, and Mingsyan Chen. 2016. Transfer Neural Trees for Heterogeneous Domain Adaptation. In *ECCV*. Springer, 399–414.
[8] Tatseng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-tao Zheng. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *CIVR*. IEEE, 48.
[9] Wenyuan Dai, Yuqiang Chen, Guirong Xue, Qiang Yang, and Yong Yu. 2008. Translated Learning: Transfer Learning across Different Feature Spaces. In *NeurIPS*. MIT Press, 353–360.
[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. IEEE, 248–255.
[11] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *ICML*, Vol. 32. ACM, 647–655.
[12] Lixin Duan, Dong Xu, and Ivor W Tsang. 2012. Learning with Augmented Features for Heterogeneous Domain Adaptation. In *ICML*. ACM, 667–674.
[13] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. 2013. Unsupervised Visual Domain Adaptation Using Subspace Alignment. In *ICCV*. IEEE, 2960–2967.
[14] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *ICML*. ACM, 1180–1189.
[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-Adversarial Training of Neural Networks. *JMLR* 17, 59 (2016), 1–35.
[16] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In *ICML*. ACM, 513–520.
[17] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. 2007. A kernel method for the two-sample-problem. In *NeurIPS*. MIT Press, 513–520.
[18] Gregory Griffin, Alex Holub, and Pietro Perona. 2007. Caltech-256 object category dataset.
[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. IEEE, 770–778.
[20] Geoffrey E Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv:1503.02531* (2015).
[21] Judy Hoffman, Erik Rodner, Jeff Donahue, Trevor Darrell, and Kate Saenko. 2013. Efficient Learning of Domain-invariant Image Representations. In *ICLR*.
[22] Judy Hoffman, Erik Rodner, Jeff Donahue, Brian Kulis, and Kate Saenko. 2014. Asymmetric and Category Invariant Feature Transformations for Domain Adaptation. *IJCV* 109, 1 (2014), 28–41.
[23] Yuanting Hsieh, Shiyen Tao, Yaohung Hubert Tsai, Yiren Yeh, and Yuchiang Frank Wang. 2016. Recognizing heterogeneous cross-domain data via generalized joint distribution adaptation. In *ICME*. IEEE, 1–6.
[24] Brian Kulis, Kate Saenko, and Trevor Darrell. 2011. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*. IEEE, 1785–1792.
[25] Jingjing Li, Ke Lu, Zi Huang, Lei Zhu, and Hengtao Shen. 2019. Heterogeneous Domain Adaptation Through Progressive Alignment. *TNNLS* 30, 5 (2019), 1381–1391.
[26] Shuang Li, Chi Harold Liu, Qiuxia Lin, Qi Wen, Limin Su, Gao Huang, and Zhengming Ding. 2020. Deep Residual Correction Network for Partial Domain Adaptation. *TPAMI* (2020), 1–1.
[27] Shuang Li, Chi Harold Liu, Binhui Xie, Limin Su, Zhengming Ding, and Gao Huang. 2019. Joint Adversarial Domain Adaptation. In *ACM MM*. ACM, 729–737.
[28] Shuang Li, Harold Chi Liu, Qiuxia Lin, Binhui Xie, Zhengming Ding, Gao Huang, and Jian Tang. 2020. Domain Conditioned Adaptation Network. In *AAAI*. AAAI Press, 11386–11393.
[29] Shuang Li, Shiji Song, Gao Huang, Zhengming Ding, and Cheng Wu. 2018. Domain Invariant and Class Discriminative Feature Learning for Visual Domain Adaptation. *TIP* 27, 9 (2018), 4260–4273.
[30] Wen Li, Lixin Duan, Dong Xu, and Ivor W Tsang. 2014. Learning With Augmented Features for Supervised and Semi-Supervised Heterogeneous Domain Adaptation.
TPAMI 36, 6 (2014), 1134–1148.
[31] Mingsheng Long, Yue Cao, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Transferable Representation Learning with Deep Adaptation Networks. *TPAMI* (2018).
[32] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. In *NeurIPS*. MIT Press, 1647–1657.
[33] Zelun Luo, Yuliang Zou, Judy Hoffman, and Fei-Fei Li. 2017. Label Efficient Learning of Transferable Representations acrosss Domains and Tasks. In *NeurIPS*. MIT Press, 165–177.
[34] Andrew L. Maas. 2013. Rectifier Nonlinearities Improve Neural Network Acoustic Models.
[35] Liqiang Nie, Wenjie Wang, Richang Hong, Meng Wang, and Qi Tian. 2019. Multimodal Dialog System: Generating Responses via Adaptive Decoders. In *ACM MM*. ACM, 1098–1106.
[36] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. 2011. Domain adaptation via transfer component analysis. *TNNLS* 22, 2 (2011), 199–210.
[37] Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *TKDE* 22, 10 (2010), 1345–1359.
[38] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. 2019. Transferrable Prototypical Networks for Unsupervised Domain Adaptation. In *CVPR*. IEEE, 2239–2247.
[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*. MIT Press, 8024–8035.
[40] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R.G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A New Approach to Cross-Modal Multimedia Retrieval. In *ACM MM*. ACM, 251–260.
[41] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. 2010. Adapting visual category models to new domains. In *ECCV*. Springer, 213–226.
[42] Xiaoxiao Shi, Qi Liu, Wei Fan, Philip S Yu, and Ruixin Zhu. 2010. Transfer Learning on Heterogenous Feature Spaces via Spectral Transformation. In *ICDE*. IEEE, 1049–1054.
[43] Xiangbo Shu, Guo-Jun Qi, Jinhui Tang, and Jingdong Wang. 2015. Weakly-Shared Deep Transfer Networks for Heterogeneous-Domain Knowledge Propagation. In *ACM MM*. ACM, 35–44.
[44] Jake Snell, Kevin Swersky, and Richard S Zemel. 2017. Prototypical Networks for Few-shot Learning. In *NeurIPS*. MIT Press, 4077–4087.
[45] Tatiana Tommasi and Tinne Tuytelaars. 2014. A testbed for cross-dataset analysis. In *ECCV*. Springer, 18–31.
[46] Yaohung Hubert Tsai, Yiren Yeh, and Yuchiang Frank Wang. 2016. Learning Cross-Domain Landmarks for Heterogeneous Domain Adaptation. In *CVPR*. IEEE, 5081–5090.
[47] Yao-Huang Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. 2016. Heterogeneous domain adaptation with label and structure consistency. In *ICASSP*. IEEE, 2842–2846.
[48] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. Simultaneous deep transfer across domains and tasks. In *ICCV*. IEEE, 4068–4076.
[49] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *CVPR*. IEEE, 2962–2971.
[50] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv:1412.3474* (2014).
[51] Chang Wang and Sridhar Mahadevan. 2011. Heterogeneous domain adaptation using manifold alignment. In *IJCAI*. Morgan Kaufmann, 1541–1546.
[52] Xinxiao Wu, Han Wang, Cuiwei Liu, and Yunde Jia. 2013. Cross-View Action Recognition over Heterogeneous Feature Spaces. In *ICCV*. IEEE, 609–616.
[53] Min Xiao and Yuhong Guo. 2015. Feature Space Independent Semi-Supervised Domain Adaptation via Kernel Matching. *TPAMI* 37, 1 (2015), 54–66.
[54] Min Xiao and Yuhong Guo. 2015. Semi-supervised subspace co-projection for multi-class heterogeneous domain adaptation. In *ECML*. Springer, 525–540.
[55] Shaoan Xie and Zibin Zheng. 2018. Learning Semantic Representations for Unsupervised Domain Adaptation. In *ICML*. ACM, 5419–5428.
[56] Yuguang Yan, Wen Li, Hanrui Wu, Huaqing Min, Mingkui Tan, and Qingyao Wu. 2018. Semi-Supervised Optimal Transport for Heterogeneous Domain Adaptation. In *IJCAI*. Morgan Kaufmann, 2969–2975.
[57] Yuan Yao, Yu Zhang, Xutao Li, and Yunming Ye. 2019. Heterogeneous Domain Adaptation via Soft Transfer Network. In *ACM MM*. ACM, 1578–1586.
[58] Han-Jia Ye, Xiang-Rong Sheng, De-Chuan Zhan, and Peng He. 2018. Distance Metric Facilitated Transportation between Heterogeneous Domains. In *IJCAI*. Morgan Kaufmann, 3012–3018.
[59] Joey Tianyi Zhou, Ivor W Tsang, Sinno Jialin Pan, and Mingkui Tan. 2014. Heterogeneous Domain Adaptation for Multiple Classes. In *AISTATS*. JMLR, 1095–1103.
[60] Junyan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *ICCV*. IEEE, 2242–2251.
[61] Junbao Zhuo, Shuhui Wang, Weigang Zhang, and Qingming Huang. 2017. Deep Unsupervised Convolutional Domain Adaptation. In *ACM MM*. ACM, 261–269.