Transferable Semantic Augmentation for Domain Adaptation

Shuang Li¹ Mixue Xie¹ Kaixiong Gong¹ Chi Harold Liu⁷ Yulin Wang² Wei Li³
¹Beijing Institute of Technology ²Tsinghua University ³Inceptio Tech.
shuangli@bit.edu.cn michellexie102@gmail.com kxgong@bit.edu.cn liuchi02@gmail.com
wang-yl19@mails.tsinghua.edu.cn liweimcc@gmail.com

Abstract

Domain adaptation has been widely explored by transferring the knowledge from a label-rich source domain to a related but unlabeled target domain. Most existing domain adaptation algorithms attend to adapting feature representations across two domains with the guidance of a shared source-supervised classifier. However, such classifier limits the generalization ability towards unlabeled target recognition. To remedy this, we propose a Transferable Semantic Augmentation (TSA) approach to enhance the classifier adaptation ability through implicitly generating source features towards target semantics. Specifically, TSA is inspired by the fact that deep feature transformation towards a certain direction can be represented as meaningful semantic altering in the original input space. Thus, source features can be augmented to effectively equip with target semantics to train a more transferable classifier. To achieve this, for each class, we first use the inter-domain feature mean difference and target intra-class feature covariance to construct a multivariate normal distribution. Then we augment source features with random directions sampled from the distribution class-wisely. Interestingly, such source augmentation is implicitly implemented through an expected transferable cross-entropy loss over the augmented source distribution, where an upper bound of the expected loss is derived and minimized, introducing negligible computational overhead. As a light-weight and general technique, TSA can be easily plugged into various domain adaptation methods, bringing remarkable improvements. Comprehensive experiments on cross-domain benchmarks validate the efficacy of TSA.

1. Introduction

Deep learning has achieved remarkable success on various vision tasks, including image recognition [16, 12, 14] and semantic segmentation [55, 49]. However, the recent success of deep learning methods heavily relies on massive labeled data. In practice, collecting abundant annotated data is expensive [30, 6, 54, 33]. Meanwhile, each domain has its own specific exploratory factors, namely semantics, e.g., the illuminations, colors, visual angles or backgrounds, resulting in the domain shift [28]. Hence, traditional deep models trained on a large dataset usually show poor generalizations on a new domain due to the domain shift issues [24, 10]. To remedy this, one appealing alternative is domain adaptation (DA), which strives to leverage the knowledge of a label-rich source domain to assist the learning in a related but unlabeled target domain [24, 10].

Prior deep DA methods can be roughly categorized as 1) statistical discrepancy minimization based methods [27, 20, 51], which leverage statistical regularizations to explicitly mitigate the cross-domain distribution discrepancy; and 2) adversarial learning based methods [21, 25, 10], which strive to learn domain-invariant representations across two domains via adversarial manners.

Indeed, these DA methods have admittedly yielded promising results, but most of them assume a shared classifier with domain-invariant representations derived. Rare attention has been paid to explicitly enhancing the adaptation ability of the source-supervised classifier, which is also fundamental to DA problems as shown in Fig. 1. To achieve classifier adaptation, Long et al. [26] introduce classifier residual learning to explicitly model the classifier difference across domains. SymNets [52] constructs three classifiers to facilitate joint distribution alignment. However, they
all rely on designing complex network architectures, which may suffer from high computational overhead and hinder the capability and versatility of these methods.

To alleviate aforementioned issues, we propose a Transferable Semantic Augmentation (TSA) approach to implicitly augment source features with target semantic guidance in the deep feature space to facilitate classifier adaptation. Specifically, TSA is motivated by the intriguing property that deep networks excel at disentangling the underlying factors of data variation and linearizing the deep features [41, 45]. There exist many different semantic transformation directions in the deep feature space, and the semantic transformation of one sample can be enforced by translating its deep feature along a certain direction, such as the direction of changing backgrounds. However, it is nontrivial to explicitly discover all kinds of semantic transformation directions. In addition, not all directions are meaningful.

Hence, to effectively explore meaningful transformation directions, we first estimate the inter-domain feature mean difference for each class as the class-wise overall semantic (i.e., an integration of various semantics in one class) bias in the deep feature space. Besides, since the specific semantic information (e.g., different backgrounds, shapes or visual angles) of source and target are different, TSA further estimates target intra-class feature covariance to effectively capture the intra-class semantic variations of target domain. To obtain the accurate estimation, we introduce a memory module to class-wisely calculate feature mean and covariance with pseudo-labeled target samples. At last, we sample semantic transformation directions for source augmentation from a multivariate normal distribution, with the estimated feature mean difference as mean and the target intra-class covariance. In this way, the overall semantic difference between domains and the target intra-class semantic variations can guide source augmented features towards target.

Furthermore, to avoid explicitly generating augmented features and improve the efficiency of TSA, we develop an expected transferable cross-entropy loss over the augmented source distribution with an upper bound derived. By minimizing the upper bound of the expected loss, the source semantic augmentation can be performed and the extra computational overhead is negligible. Then, the trained source classifier can be successfully adapted to target.

Contributions of this work are summarized as follows:

- We propose a novel Transferable Semantic Augmentation (TSA) method for classifier adaptation, which enables source feature augmentation towards target in an implicit manner. Notably, TSA introduces no extra network modules over the backbone network, making it simple to implement and computationally efficient.

- We develop a novel expected transferable cross-entropy loss over the augmented source distribution for DA, which greatly enhances the classifier adaptation ability. Moreover, as a light-weight and general technique, TSA can be easily plugged into various DA methods to significantly boost their performances.

- Extensive experiments on several cross-domain benchmarks, including Office-Home, Office-31, VisDA-2017 and digits demonstrate that TSA can consistently yield significant performance improvements.

2. Related Work

2.1. Feature Adaptation

Existing domain-invariant feature learning methods mainly fall into two categories. One is the discrepancy-based works [24, 27, 20, 51, 28, 31, 7, 19], which concentrates on mitigating the domain difference by minimizing some statistical discrepancy metrics. To name a few, DANN [24] minimizes maximum mean discrepancy (MMD) on task-specific layers to close the domain gap. Further, JAN in [27] introduces a joint MMD to enforce joint distribution alignment between domains. [51] proposes margin disparity discrepancy (MDD) to facilitate precise alignment with theoretical guarantees. Based on the optimal transport (OT) distance [53], [19] proposes an enhanced transport distance (ETD) to learn the optimal transport plan.

Another category is adversarial-based works [39, 44, 10, 25, 21], which learns domain-invariant features in adversarial manners. For example, DANN [10] introduces a domain discriminator to force the features of two domains indistinguishable. Following DANN, CDAN [25] integrates classification information into the domain discriminator to alleviate class-wise mismatching. While MCD [39] employs a new adversarial paradigm where the adversarial manner occurs between the feature extractor and classifiers rather than the feature extractor and the domain discriminator.

In contrast, instead of learning domain-invariant features, TSA harnesses the augmented source features to ameliorate the generalization ability of source classifier, which is also crucial for tackling DA problems.

2.2. Classifier Adaptation

On par with feature adaptation methods, classifier adaptation is also an indispensable part of DA, since the assumption that source and target domains can share one identical classifier is rather restrictive in practical scenarios.

A branch of classifier adaptation methods augments training samples towards target style to adapt the classifier [5, 2, 13, 43, 22, 23, 8]. To be specific, Hoffman et al. [13] exploit pixel cycle-consistency to make transformed images visually close to target domain images and use transformed samples to train the target model. UNIT [23] exploits image-to-image translation framework based
3. Transferable Semantic Augmentation

3.1. Motivation and Preliminaries

This work delves into the classifier adaptation for DA problems, which has not yet been fully explored. We propose a Transferable Semantic Augmentation (TSA) approach to effectively augment source features towards target semantics (e.g., backgrounds, view angles or colors) by generating a very large (infinite) number of meaningful semantic transformation directions. The augmented source features will facilitate adapting the classifier from source to target successfully. TSA is inspired by the fact that deep neural networks surprisingly excel at linearizing features of input data, making it linearly separable [41, 45]. Therefore, the relative positions in deep feature space can represent certain semantic transformations in the original input space. Based on this observation, MCF [9] and ISDA [45] have performed semantic augmentation for linear models and deep networks, respectively. However, neither of them considers the crucial domain shift issue in the real world.

To tackle the cross-domain semantic mismatch problems, we need to perform meaningful and target-wise semantic augmentations for source data. In other words, the semantic augmentation directions should effectively guide source features to be target style, such as the shape, color, visual angle or background of target data.

Formally, we denote the source and target domains in DA as $\mathcal{S} = \{(x_{si}, y_{si})\}_{i=1}^{n_s}$ with $n_s$ labeled samples and $\mathcal{T} = \{x_{tj}\}_{j=1}^{n_t}$ with $n_t$ unlabeled samples, respectively, where $y_{si} \in \{1, 2, ..., C\}$ is the label of $x_{si}$. Note that, source and target data are sampled from different distributions, such that certain discrepancy may exist between their corresponding ideal classifiers. Our goal is to adapt the source-supervised classifier to target by augmenting the source feature representation $f_s = F(x_s, \Theta_F) \in \mathbb{R}^K$, where $K$ is the number of feature embedding dimensions and $F(\cdot)$ is the feature extractor parameterized by $\Theta_F$.

3.2. Cross-Domain Semantic Augmentation

As aforementioned, certain translating directions in deep feature space represent meaningful semantic transformations in the original input space. Thus, we attempt to augment source data towards target in deep feature space. However, manually searching for such cross-domain transformation directions is nontrivial. To address the issue, for each class, we propose to sample random vectors from a multivariate normal distribution with inter-domain feature mean difference as mean and class-conditional target covariance as the covariance matrix. By doing so, rich meaningful cross-domain transformation directions can be discovered. Then, we design an expected transferable cross-entropy loss over the augmented feature distribution to effectively and efficiently adapt the classifier from source to target domain.

3.2.1 Class-wise Overall Semantic Guidance.

Due to the domain shift, for the same class, the means of source and target samples are different in the deep feature space. Such difference reflects the bias of the overall se-
mantics, an integration of various semantics in one class. To conduct useful transformations, we force the mean alignment between augmented source features and target features class-wisely, which effectively bridges the overall semantic bias across two domains. Specifically, for each class \( c \), let \( \mu^s_c \) and \( \mu^t_c \) denote the estimated mean feature vector for source and target domains, respectively. Here, pseudo labels for target data are employed as \( y'_{tj} = \arg \max_c P_{tj}^c \) to address the label scarce problem in target domain, where \( P_{tj} \) is the softmax outputs of target sample \( x_{tj} \). Then, the inter-domain mean difference \( \Delta \mu^c = \mu^t_c - \mu^s_c \) can be exploited to mitigate such overall semantic bias.

### 3.2.2 Semantic Transformation Direction Learning.

To facilitate meaningful cross-domain semantic augmentations, all possible feature transformation directions towards target style need to be discovered. However, only utilizing the inter-domain mean difference is infeasible to fully explore target semantic variations.

Therefore, to address this problem, we further exploit the intra-class covariance of feature to capture target semantic variations. Specifically, we randomly sample semantic transformation directions from a multivariate distribution \( N(\Delta \mu^c, \Sigma^c_i) \) for each class \( c \), and \( \Sigma^c_i \) is the corresponding target intra-class covariance, which contains rich target semantic variations. Here, we develop the normal distribution class-wisely, since the semantic knowledge in each class may differ from each other vastly. Notably, \( \Delta \mu^c \) attends to mitigating the overall semantic bias for class \( c \), and \( \Sigma^c_i \) focuses on providing abundant target intra-class semantic variation knowledge. By adding the sampled transformation vectors, the augmented source features will be close to target domain and vary along target semantic variations.

### 3.2.3 Memory Module.

To efficiently implement TSA in an end-to-end training manner, for each class \( c \), we propose to estimate \( \Delta \mu^c \) and \( \Sigma^c_i \) according to a memory module that stores all the latest features. By contrast, ISDA \cite{45} proposes an iterative manner by accumulating statistics of batches from first to current batch. However, the weight distribution of network in early stage will vastly differ from that in later stage. Thus, due to its accumulation property, such iterative manner may encounter the issue that out-of-date features will bias the estimation of mean and covariance. To avoid this, we use a memory module to cache all features of two domains and update in each batch. By doing so, we can discard those out-of-date features and replace them with the latest ones at the negligible cost of memory. Formally, in each iteration \( i \), we will update a batch of features and corresponding labels in memory module \( \mathbb{M} \):

\[
\begin{aligned}
f^M_j &\leftarrow f^{(i)}_j, y^M_j &\leftarrow y^{(i)}_j, & j \in \mathbb{B}^{(i)},
\end{aligned}
\]

where \( j \) is the sample index within a batch \( \mathbb{B}^{(i)} \) and \( f^M_j, y^M_j \) is the feature/label stored in memory module \( \mathbb{M} \). Comparisons of the two manners will be shown in the experiment.

### 3.2.4 Sampling Strategy.

In the early training stage, the predictions of target samples are inevitably not that accurate. Thus, the estimation of target mean and covariance may not be quite informative as expected. To alleviate this issue, for each class \( c \), we sample semantic transformation directions from \( N(\lambda \Delta \mu^c, \lambda \Sigma^c_i) \), where \( \lambda \) is a positive parameter to control the augmentation strength. Since the target predictions will become more and more accurate as the training progresses, here we set \( \lambda = (t/T) \times \lambda_0 \), where \( t \) and \( T \) are the current and maximum iterations respectively, and \( \lambda_0 \) is a hyper-parameter. Consequently, as the training goes on, \( \lambda \) will gradually grow from 0 to \( \lambda_0 \). And this sampling strategy will reduce the impact of less accurate estimations of mean and covariance at the beginning of the training stage.

### 3.3 Transferable Cross-Entropy Loss Learning

Once \( C \) sampling distributions are constructed, each source deep feature \( f_{si} \) can conduct various semantic transformations along the random directions sampled from \( N(\lambda \Delta \mu^c, \lambda \Sigma^c_i) \) to generate the augmented feature \( f_{si}^c \), i.e., \( f_{si} \sim N(f_{si} + \lambda \Delta \mu^c, \lambda \Sigma^c_i) \). Considering a naive method, we can augment each \( f_{si} \) for \( M \) times with its label preserved, which will result in an augmented feature set \( \{ (f_{s1}^1, y_{si}), (f_{s1}^2, y_{si}), ..., (f_{s1}^M, y_{si}) \}^{n_s=1} \). Based on this, the source network can be trained with a traditional cross-entropy loss on the augmented feature set:

\[
\mathcal{L}_M(\theta_F, \mathbf{W}, b) = \frac{1}{n_s} \sum_{i=1}^{n_s} \frac{1}{M} \sum_{m=1}^{M} - \log \left( \frac{w^T_{ei} f_{si}^m + b_{yi}}{\sum_{i=1}^{n_s} w^T_{ei} f_{si}^m + b_{yi}} \right),
\]

where \( \mathbf{W} = [w_1, w_2, ..., w_C]^T \in \mathbb{R}^{C \times K} \) and \( b = [b_1, b_2, ..., b_C]^T \in \mathbb{R}^C \) are the weight matrix and bias vector of the last fully connected layer, respectively.

To achieve desired performance, \( M \) is usually large, resulting in unexpected computational cost. In TSA, instead of explicitly generating the augmented source features for \( M \) times, we intend to implicitly generate infinite augmented source features. When \( M \) approaches infinity, we can derive an upper-bound loss according to the Law of Large Numbers over the augmented source distribution.

Specifically, in the case of \( M \to \infty \), our proposed expected transferable cross-entropy loss over the augmented source feature distribution is defined as:

\[
\lim_{M \to \infty} \mathcal{L}_M = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbb{E}_{\mathbb{F}_{si}} \left[ - \log \left( \frac{w^T_{ei} x_{si}^c + b_{ ei}}{\sum_{i=1}^{n_s} w^T_{ei} x_{si}^c + b_{ ei}} \right) \right] = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbb{E}_{\mathbb{F}_{si}} \left[ \log \left( \sum_{c=1}^{C} (w^T_{ei} - w^T_{ei} y_{si} + b_{ei} - b_{ei}) \right) \right].
\]
However, it is infeasible to tackle Eq (3) directly. According to the Jensen’s inequality [3] and the concave property of logarithmic function $\log(\cdot)$, we have $E[\log(X)] \leq \log(E[X])$. Thus, we can derive the upper bound of the expected loss as follows:

\[
\lim_{M \to \infty} \mathcal{L}_M \leq \frac{1}{n_s} \sum_{i=1}^{n_s} \log \left( \sum_{c=1}^{C} \frac{c^{\top} (w_c^T - w_{c, i}^T) f_{si} + (b_c - b_{y_{si}})}{C} \right) = \frac{1}{n_s} \sum_{i=1}^{n_s} \log \left( \sum_{c=1}^{C} \frac{c^{\top} (w_c^T - w_{c, i}^T) f_{si} + (b_c - b_{y_{si}})}{C} \right).
\]

(4)

Due to $f_{si} \sim N(f_{si} + \lambda \Delta \mu_{ys_i}, \lambda \Sigma_{ys_i})$, we can derive that $(w_c^T - w_{c, i}^T)f_{si} + (b_c - b_{y_{si}}) \sim N((w_c^T - w_{c, i}^T) + (b_c - b_{y_{si}}), \sigma_{si}^2)$, where $\sigma_{si}^2 = \lambda (w_c^T - w_{c, i}^T) \Sigma_{ys_i} (w_c - w_{c, i})$. Leveraging the moment-generating function $E[e^{\alpha X}] = e^{\alpha \mu + \frac{1}{2} \alpha^2 \sigma^2}$, $X \sim N(\mu, \sigma)$, we can obtain

\[
\lim_{M \to \infty} \mathcal{L}_M \leq \mathcal{L}_\infty = -\frac{1}{n_s} \sum_{i=1}^{n_s} \log \left( \frac{c^{\top} \tilde{y}_{si}}{C} \right),
\]

(5)

where $\tilde{y}_{si} = \tilde{y}_{si}^c + \lambda (w_c^T - w_{c, i}^T) \Delta \mu_{ys_i} + \sigma_{si}^2$ and $\tilde{y}_{si}^c$ is the $c$-th element of logits output of $x_{si}$.

Essentially, Eq (5) gives us a surrogate loss $\mathcal{L}_\infty$ for optimizing the expected transferable cross-entropy loss, which could effectively adapt the source classifier to target with the augmented source features. Moreover, the proposed novel loss allows us to utilize TSA as a plug-in module for other DA methods to further improve their transferability.

**Discussion.** Although the derivation of the upper bound is kind of similar with ISDA [45], the motivation of our method is essentially different. We aim to enhance the classifier adaptation for DA where target domain is unlabeled and large domain shift exists between domains. Thus, the proposed $\Delta \mu^c$ and $\Sigma^c_i$ are crucial to successfully generate transformation directions that can really bridge the domain gap. While ISDA is a completely supervised algorithm and ignores the crucial domain shift issues in reality. In the ablation study, we show that the performance of directly applying ISDA in DA context is inferior. Besides, different from the iterative estimation manner in ISDA, our memory module manner can achieve more accurate estimations of the mean and covariance, which is shown in the experiment.

### 3.4. Overall Formulation

In information theory, mutual information $I(X; Y)$ measures how related two random variables $X$ and $Y$ are. Actually, strong correlations between target features and predictions will benefit our semantic augmentations, because the extracted features will be more informative and contain more important semantics for predictions, ignoring trivial semantics. Thus, we employ the mutual information maximization on target data, i.e., minimizing the loss in Eq (6).

\[
\mathcal{L}_{MI} = \frac{1}{n_t} \sum_{j=1}^{n_t} \log \frac{n_t}{n_s} \sum_{c=1}^{C} P'^c_i \log P'^c_i,
\]

(6)

where $P' = \frac{1}{n_t} \sum_{j=1}^{n_t} P_{ij}$. Since target domain is unlabeled, we use the average of target predictions to approximate the ground-truth distribution on target domain.

To sum up, the overall objective function of TSA is:

\[
\mathcal{L}_{TSA} = \mathcal{L}_\infty + \beta \mathcal{L}_{MI},
\]

(7)

where $\beta$ is a trade-off parameter. The effects of different parts in TSA will be analyzed in details in ablation study.

### 3.5. Theoretical Insight

To theoretically understand TSA, we introduce the domain adaptation theory proposed by [11], which reveals the ingredients of target generalization error $\epsilon_t$. Formally, let $H$ denote the hypothesis space and $h \in H$ denote the classifier, we can formulate the upper bound of $\epsilon_t$ as:

\[
\epsilon_t(h) \leq \epsilon_s(h) + \frac{1}{2} d_{H\Delta H}(S, T) + \lambda^*, \forall h \in H,
\]

(8)

where $\epsilon_s(h)$ is the source generalization error of $h$, $d_{H\Delta H}(S, T)$ is the $H\Delta H$-distance between source and target domains, and $\lambda^* = \epsilon_s(h^*) + \epsilon_t(h^*)$ denotes the error of an ideal joint hypothesis $h^*$ on source and target domains.

With the supervision of labeled source data, $\epsilon_s(h)$ is well bounded. Besides, TSA generates transformation directions from the constructed multivariate distribution to augment source features towards target domain. Since these augmented source features can fill the domain gap, TSA further tightens the $d_{H\Delta H}(S, T)$. Moreover, TSA implicitly generates infinite augmented source features that are close to target domain class-wisely, enabling the classifier to jointly minimize $\epsilon_s(h^*)$ and $\epsilon_t(h^*)$ of the shared error $\lambda^*$ on the augmented training set. To sum up, TSA complies with the theory, thus further enhancing the transferability.

### 4. Experiment

#### 4.1. Datasets

**Office-Home** [42] is a challenging benchmark for domain adaptation, which contains 15,500 images in 65 classes drawn from 4 distinct domains: Artistic (Ar), Clip Art (Cl), Product (Pr), and Real-World (Rw).

**Office-31** [38] is a classical cross-domain benchmark, including images in 31 classes drawn from 3 distinct domains: Amazon (A), Webcam (W) and DSLR (D).

**VisDA-2017** [35] is a large-scale dataset for visual domain adaptation, containing over 280K images across 12 categories. Following [39], we choose the synthetic domain with 152,397 images as source and the realistic domain with 55,388 images as target.
Table 1. Accuracy (%) on Office-Home for UDA (ResNet-50).

<table>
<thead>
<tr>
<th>Method</th>
<th>Ar→Cl</th>
<th>Ar→Pr</th>
<th>Ar→Rw</th>
<th>Cl→Ar</th>
<th>Cl→Pr</th>
<th>Cl→Rw</th>
<th>Pr→Ar</th>
<th>Pr→Cl</th>
<th>Pr→Rw</th>
<th>Rw→Ar</th>
<th>Rw→Cl</th>
<th>Rw→Pr</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAN [27]</td>
<td>45.9</td>
<td>61.2</td>
<td>68.9</td>
<td>50.4</td>
<td>59.7</td>
<td>61.0</td>
<td>45.8</td>
<td>43.4</td>
<td>70.3</td>
<td>63.9</td>
<td>52.4</td>
<td>76.8</td>
<td>58.3</td>
</tr>
<tr>
<td>TAT [22]</td>
<td>51.6</td>
<td>69.5</td>
<td>75.4</td>
<td>59.4</td>
<td>69.5</td>
<td>68.6</td>
<td>59.5</td>
<td>50.5</td>
<td>76.8</td>
<td>70.9</td>
<td>56.6</td>
<td>81.6</td>
<td>65.8</td>
</tr>
<tr>
<td>TPN [31]</td>
<td>51.2</td>
<td>71.2</td>
<td>76.0</td>
<td>65.1</td>
<td>72.9</td>
<td>72.8</td>
<td>55.4</td>
<td>48.9</td>
<td>76.5</td>
<td>70.9</td>
<td>53.4</td>
<td>80.4</td>
<td>66.2</td>
</tr>
<tr>
<td>ETD [19]</td>
<td>51.3</td>
<td>71.9</td>
<td>85.7</td>
<td>57.6</td>
<td>69.2</td>
<td>73.7</td>
<td>57.8</td>
<td>51.2</td>
<td>79.3</td>
<td>70.2</td>
<td>57.5</td>
<td>82.1</td>
<td>67.3</td>
</tr>
<tr>
<td>SymNets [52]</td>
<td>47.7</td>
<td>72.9</td>
<td>78.5</td>
<td>64.2</td>
<td>71.3</td>
<td>74.2</td>
<td>64.2</td>
<td>48.8</td>
<td>79.5</td>
<td>74.5</td>
<td>52.6</td>
<td>82.7</td>
<td>67.6</td>
</tr>
<tr>
<td>BNM [7]</td>
<td>52.3</td>
<td>73.9</td>
<td>80.0</td>
<td>63.3</td>
<td>72.9</td>
<td>74.9</td>
<td>61.7</td>
<td>49.5</td>
<td>79.7</td>
<td>70.5</td>
<td>53.6</td>
<td>82.2</td>
<td>67.9</td>
</tr>
<tr>
<td>MDD [51]</td>
<td>54.9</td>
<td>73.7</td>
<td>77.8</td>
<td>60.0</td>
<td>71.4</td>
<td>71.8</td>
<td>61.2</td>
<td>53.6</td>
<td>78.1</td>
<td>72.5</td>
<td>60.2</td>
<td>82.3</td>
<td>68.1</td>
</tr>
<tr>
<td>GSP [47]</td>
<td>56.8</td>
<td>75.5</td>
<td>78.9</td>
<td>61.3</td>
<td>69.4</td>
<td>74.9</td>
<td>61.3</td>
<td>52.6</td>
<td>79.9</td>
<td>73.3</td>
<td>54.2</td>
<td>83.2</td>
<td>68.4</td>
</tr>
<tr>
<td>GVB-GD [8]</td>
<td>57.0</td>
<td>74.7</td>
<td>79.8</td>
<td>64.6</td>
<td>74.1</td>
<td>74.6</td>
<td>65.2</td>
<td>55.1</td>
<td>81.0</td>
<td>74.6</td>
<td>59.7</td>
<td>84.3</td>
<td>70.4</td>
</tr>
<tr>
<td>ResNet-50 [12] +TSA</td>
<td>53.6</td>
<td>75.1</td>
<td>80.0</td>
<td>64.4</td>
<td>73.6</td>
<td>72.5</td>
<td>62.3</td>
<td>49.4</td>
<td>77.5</td>
<td>72.2</td>
<td>58.8</td>
<td>82.1</td>
<td>68.3</td>
</tr>
<tr>
<td>DANN [10] +TSA</td>
<td>45.6</td>
<td>59.3</td>
<td>70.1</td>
<td>47.0</td>
<td>59.5</td>
<td>60.9</td>
<td>46.1</td>
<td>43.7</td>
<td>68.5</td>
<td>63.2</td>
<td>51.8</td>
<td>76.8</td>
<td>57.6</td>
</tr>
<tr>
<td>CDAN [25] +TSA</td>
<td>50.7</td>
<td>70.6</td>
<td>76.0</td>
<td>57.6</td>
<td>70.0</td>
<td>70.0</td>
<td>57.4</td>
<td>50.9</td>
<td>77.3</td>
<td>70.9</td>
<td>56.7</td>
<td>81.6</td>
<td>65.8</td>
</tr>
<tr>
<td>BSP [4] +TSA</td>
<td>52.0</td>
<td>68.6</td>
<td>76.1</td>
<td>58.0</td>
<td>70.3</td>
<td>70.2</td>
<td>58.6</td>
<td>50.2</td>
<td>77.6</td>
<td>72.2</td>
<td>59.3</td>
<td>81.9</td>
<td>66.3</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>57.6</strong></td>
<td><strong>75.8</strong></td>
<td><strong>80.7</strong></td>
<td><strong>64.3</strong></td>
<td><strong>76.3</strong></td>
<td><strong>75.1</strong></td>
<td><strong>66.7</strong></td>
<td><strong>55.7</strong></td>
<td><strong>81.2</strong></td>
<td><strong>75.7</strong></td>
<td><strong>61.9</strong></td>
<td><strong>83.8</strong></td>
<td><strong>71.2</strong></td>
</tr>
</tbody>
</table>

Digital Datasets contain 4 standard digital datasets: MNIST [17], USPS [15], Street View House Numbers (SVHN) [29] and synthetic digits dataset (SYN) [11]. All of these datasets provide number images from 0 to 9. We construct four transfer tasks: MNIST to USPS (M → U), USPS to MNIST (U → M), SVHN to MNIST (SV → M) and SYN to MNIST (SY → M).

4.2. Implementation Details

For a fair comparison, we use ResNet [12] pre-trained on ImageNet [37] as the backbone network for datasets: Office-Home, Office-31 and VisDA-2017. As for digital datasets, we employ the same network structures in [21] and train the networks from scratch. In this paper, all experiments are implemented via PyTorch [32]. We adopt mini-batch SGD optimizer with momentum 0.9 for network optimization, and deep embedded validation [50] to select hyper-parameters $\lambda_0$ from $\{0.1, 0.25, 0.5, 0.75, 1.0\}$ and $\beta$ from $\{0.01, 0.05, 0.1, 0.15, 0.2\}$, and we found $\lambda_0 = 0.25, \beta = 0.1$ works well on all datasets.

We evaluate our approach by applying TSA to the source-only model and several mainstream DA methods, i.e., DANN [10], CDAN [25], BSP [4] and JADA [21], based on their open-source codes. For each task, we report the average accuracy of 3 random trials. Code of TSA is available at https://github.com/BIT-DA/TSA.

4.3. Results

Results on Office-Home are reported in Table 1. Office-Home is a challenging dataset with large domain discrepancy for DA. TSA consistently improves the generalization ability of the four based methods. Specifically, TSA empowers BSP with 4.9% improvement, achieving the highest accuracy 71.2% on average. Based on these promising results, we can infer that TSA can stably enhance the transferability of classifiers on this difficult cross-domain dataset.

Results on Office-31 are presented in Table 2. Compared with feature adaptation methods (e.g., JAN and DANN), ResNet-50+TSA surpasses them by a large margin, indicating that classifier adaptation is also indispensable for DA. In particular, TSA brings a large improvement of 7.2% to DANN, revealing that TSA is complementary to previous DA methods. Besides, ResNet-50+TSA also surpasses recent classifier adaptation methods, e.g., SymNets and JADA, indicating that classifier adaptation is also indispensable for DA. In particular, TSA brings a large improvement of 7.2% to DANN, revealing that TSA is complementary to previous DA methods.
forms other augmentation methods, like DMRL [46] and DM-ADA [48]. This is mainly due to the fact that TSA exploits the mean difference and target covariance to fully capture the meaningful semantic information class-wisely, achieving better augmentation results. Besides, TSA also brings remarkable improvements to other baseline methods, which validates the effectiveness and versatility of TSA.

**Results on Digital Datasets** are reported in Table 4. It is noteworthy that DIFA can be seen as DANN combined with extra GAN-based modules which are designed for generating training samples. Thus, it is fair to compare DIFA with DANN+TSA. DANN+TSA exceeds DIFA on average and especially surpasses it by a large margin on U → M. Besides, TSA does not require to introduce any auxiliary networks, which is extremely easy to implement. For CDAN and JADA, TSA also boosts their performance on average. These encouraging results validate that our semantic augmentation approach can derive a robust classifier efficiently.

### 4.4. Analysis

**Ablation Study.** For each class $c$, the multivariate normal distribution for sampling transformation directions consists of two important parts: 1) inter-domain feature mean difference $\Delta \mu^c$ and 2) target intra-class covariance matrix $\Sigma_{c}^t$. Only using $\Sigma_{c}^t$ to build sampling distribution $N(0, \lambda \Sigma_{c}^t)$ will suffer from overall semantic bias due to the natural dissimilarity between two domains. While only using $\Delta \mu^c$ to construct the distribution $N(\lambda \Delta \mu^c, 0)$ cannot fully capture semantic variations underlying target domain.

Therefore, to investigate the effects of different components in the sampling distribution, we conduct extensive ablation studies of TSA based on DANN [10] in Table 5. The results manifest that our strategy achieves the best performance, indicating that our sampling distribution can yield meaningful transformation directions to guide the augmented source features towards target.

Fixing the sampling distribution $N(\lambda \Delta \mu^c, \lambda \Sigma_{c}^t)$, we further explore the impacts of mutual information maxi-
mization loss $\mathcal{L}_{MI}$ in TSA. In Table 5, TSA seems to be more effective when incorporating $\mathcal{L}_{MI}$, since mutual information maximization helps achieve more accurate target predictions and more informative target features, which is beneficial for capturing the underlying target semantics. Besides, DANN+TSA (w/o $\mathcal{L}_{MI}$) achieves much better results than DANN+ISDA. The reason is that ISDA [45] does not consider the domain shift problem and fails to capture target semantics, resulting in poor classifier adaptability.

**Visualization for Source Augmentation.** To intuitively validate that TSA can generate meaningful augmented features, we design a reverse mapping algorithm (presented in the supplement) to search images corresponding to the augmented features. From Fig. 3, we can observe that TSA can alter the semantics of source images, such as backgrounds, colors and shapes of numbers. Moreover, some augmented images closely resemble target images. This verifies that TSA can indeed generate diverse, meaningful and even target-specific augmented features. More augmented images are shown in the supplement.

**Inter Twinning Moons 2D Problems.** The goal of TSA is to learn transferable classifiers. We carry out experiments on inter twinning moons 2D problem [34] to illustrate the adaptation of classification boundaries brought by TSA. In Fig. 4, source-only model and DANN correctly classify almost all source samples, but misclassify many target samples. By contrast, the classification boundaries of source-only+TSA and DANN+TSA are adapted to target data clearly. The apparent classifier movement reveals that TSA can enhance the classifier transferability notably.

**Estimation Bias Comparisons.** In this section, we compare the estimation bias of the iterative manner in ISDA [45] and our memory module manner. Specifically, in each epoch, we calculate the average Euclidean distances between ideal and practical estimations of the mean and covariance for all $C$ classes, respectively. The ideal estimation denotes using all features that are freshly produced by feeding the whole dataset into current network to estimate $\Delta \mu^c$ and $\Sigma^c_t$ for each class $c$; the practical estimation denotes leveraging the iterative manner in ISDA or our memory module manner to class-wisely calculate $\Delta \mu^c$ and $\Sigma^c_t$. The results are shown in Fig. 5. We observe that the mean and covariance estimations by our memory module manner better approximate the ideal estimation. Based on the results, we can infer that our manner better contributes to deriving more precise mean and covariance estimations.

**Hyper-parameter Sensitivity.** Hyper-parameter $\lambda_0$ controls the transformation strength and $\beta$ balances the two losses in TSA. To check the parameter sensitivity of TSA, we conduct experiments on two random tasks Ar$\rightarrow$Pr and Cl$\rightarrow$Rw (Office-Home) based on CDAN.

**5. Conclusion**

This paper presents a transferable semantic augmentation (TSA) approach to ameliorate the adaptation ability of the classifier by optimizing a novel transferable loss over the implicitly augmented source distribution, introducing negligible computational costs. TSA is applicable to various DA methods and can yield significant improvements. Comprehensive experiments on several cross-domain datasets have demonstrated the efficacy and versatility of TSA.
References


[8] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In CVPR, June 2020. 2, 3, 6


[20] Shuang Li, Chi Harold Liu, Quxia Lin, Qi Wen, Limin Su, Gao Huang, and Zhengming Ding. Deep residual correction network for partial domain adaptation. TPAMI, pages 1–1, 2020. 1, 2


