Unsupervised Domain Adaptation via Structurally Regularized Deep Clustering

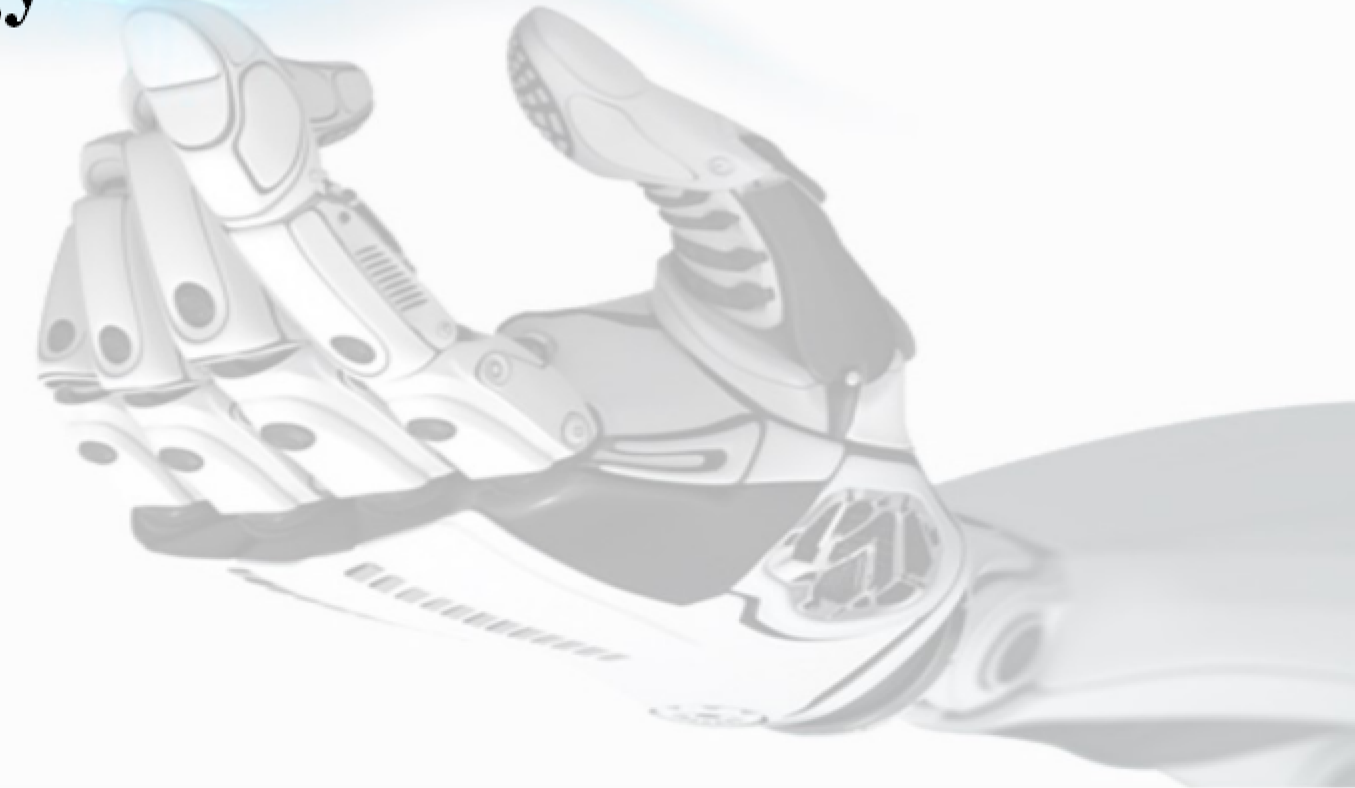
Authors: Hui Tang, Ke Chen, Kui Jia



Institution: South China University of Technology

[1] Hui Tang, Ke Chen, Kui Jia. Unsupervised Domain Adaptation via Structurally Regularized Deep Clustering. CVPR2020.

CONTENTS



PROBLEM DEFINITION

- Source domain $S = \{(x_j^s, y_j^s)\}_{j=1}^{n_s}$
- Target domain $T = \{x_i^t\}_{i=1}^{n_t}$

A shared label space Y: $y^s, y^t \in \{1,2,\ldots,K\}$.

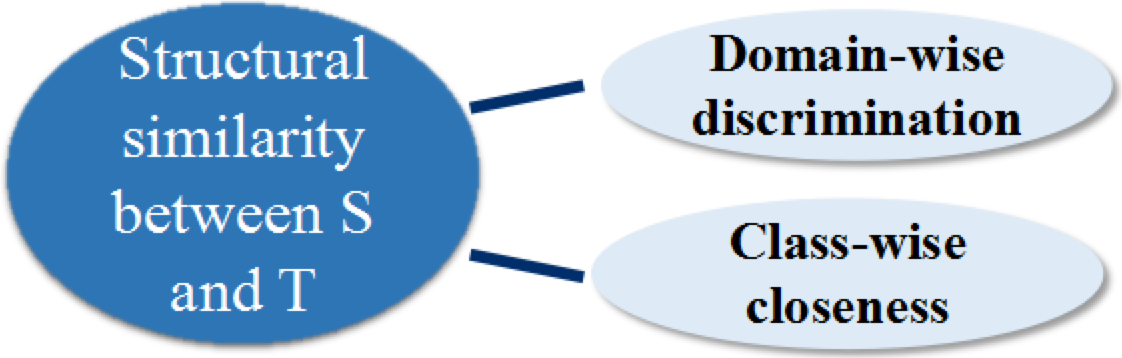
- Feature embedding function $\varphi(\cdot;\theta): X \to Z$ lifts any $x \in X$ to the feature space Z, i.e. $z = \varphi(x)$.
- Classifier $f(\cdot;\vartheta): Z \to R^K$ with softmax at the top outputs a probability vector $p = \mathrm{softmax}(f(z))$.

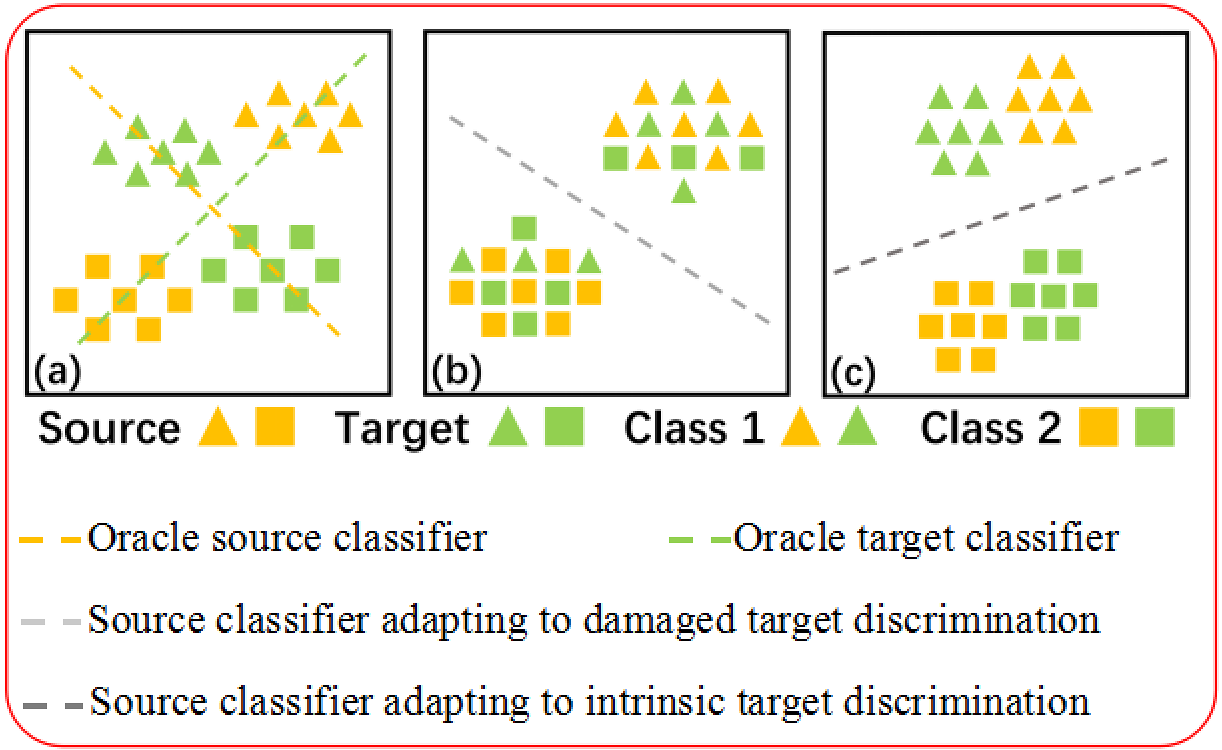
➢ **Objective: Given labeled data on S, UDA is to predict class labels for unlabeled data sampled from T by learning $\varphi(\cdot)$ and $f(\cdot)$ on both $\{(x_j^s, y_j^s)\}_{j=1}^{n_s}$ and $\{x_i^t\}_{i=1}^{n_t}$.**

- Transductive UDA is to measure performance of the learned $\varphi(\cdot)$ and $f(\cdot)$ on $\{x_i^t\}_{i=1}^{n_t}$.
- Inductive UDA is to evaluate on held-out instances sampled from the same T.
- This subtle difference is in fact important since off-the-shelf models are expected.

OUR UNCOVERING STRATEGY

- Assumption of structural domain similarity (a):





- Existing transferring strategy of learning aligned features across domains (b) → a sub-optimal generalization.

- Based on the assumption, our strategy of uncovering intrinsic discrimination of target data (c) → an adapted classifier closer to oracle target classifier.

OUR UNCOVERING STRATEGY

- Motivation: Mainstream UDA methods take the transferring strategy, which has a potential risk of damaging the intrinsic discrimination of target data, as discussed in [2, 3, 4].
- Solution: We directly uncover the intrinsic target discrimination via discriminative clustering of target data and constrain the clustering solutions using structural source regularization that hinges on our assumed structural domain similarity.

[2] Xinyang Chen, Sinan Wang, Mingsheng Long, et al. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. ICML2019.
[3] Yuan Shi and Fei Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. ICML2012.
[4] Han Zhao, Remi Tachet Des Combes, Kun Zhang, et al. On learning invariant representations for domain adaptation. ICML2019.

OUR UNCOVERING STRATEGY

➢ **Structurally Regularized Deep Clustering (SRDC)** *implicitly* achieves feature alignment between the two domains.

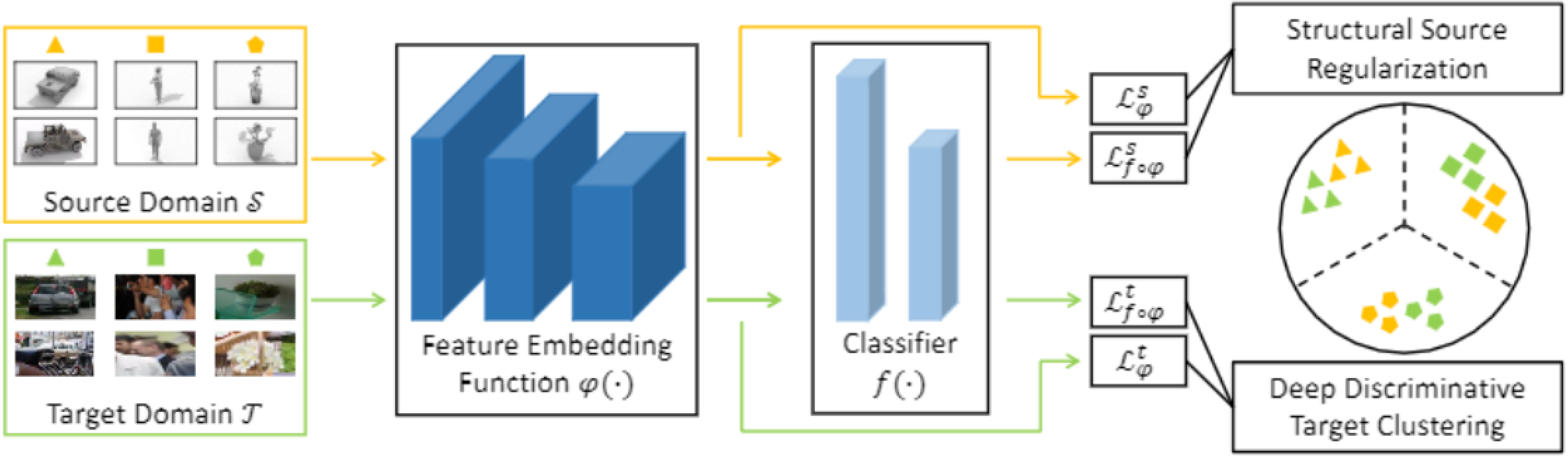


Figure 1. (Best viewed in color). Pipeline of our proposed method. The colors of orange and green denote different domains. The shapes of triangle, rectangle, and pentagon denote different classes. Based on the assumed structural similarity across domains, our losses of $\mathcal{L}^t_{f\circ\varphi}$ and $\mathcal{L}^t_{\varphi}$ uncover the intrinsic discrimination of unlabeled target data and those of $\mathcal{L}^s_{f\circ\varphi}$ and $\mathcal{L}^s_{\varphi}$ transfer the global, discriminative structure of labeled source data via joint training. An example of the effect of our proposed method is shown in the circle.

OUR UNCOVERING STRATEGY

- **Deep discriminative target clustering** [5] minimizes the KL divergence between predictive label distribution of the network and an introduced auxiliary one:

$$\min_{Q^t, \{\theta, \vartheta\}} \mathcal{L}^t_{f \circ \varphi} = \mathrm{KL}(Q^t || P^t) + \sum_{k=1}^{K} \varrho^t_k \log \varrho^t_k, \qquad (1)$$

We collectively write $\{p_i^t\}_{i=1}^{n_t}$ as P^t.

Q^t is the introduced auxiliary counterpart.

$\rho_k^t = 1/n_t \sum_{i=1}^{n_t} q_{i,k}^t$ is the predicted target label distribution.

- Optimization takes alternating steps:

Auxiliary distribution update

$$q^t_{i,k} = \frac{p^t_{i,k} / (\sum_{i'=1}^{n_t} p^t_{i',k})^{\frac{1}{2}}}{\sum_{k'=1}^{K} p^t_{i,k'} / (\sum_{i'=1}^{n_t} p^t_{i',k'})^{\frac{1}{2}}}. \qquad (2)$$

Network update

$$\min_{\theta, \vartheta} -\frac{1}{n_t} \sum_{i=1}^{n_t} \sum_{k=1}^{K} q^t_{i,k} \log p^t_{i,k}. \qquad (3)$$

- We also enhance target discrimination with deep embedding clustering in the feature space [6]:

$$\tilde{p}^t_{i,k} = \frac{\exp((1 + ||z^t_i - \mu_k||^2)^{-1})}{\sum_{k'=1}^{K} \exp((1 + ||z^t_i - \mu_{k'}||^2)^{-1})}. \qquad (4)$$

$$\min_{\tilde{Q}^t, \theta, \{\mu^t_k\}_{k=1}^{K}} \mathcal{L}^t_{\varphi} = \mathrm{KL}(\tilde{Q}^t || \tilde{P}^t) + \sum_{k=1}^{K} \tilde{\varrho}^t_k \log \tilde{\varrho}^t_k, \qquad (5)$$

where μ_k is the learnable cluster center and $\tilde{p}^t_{i,k}$ is a probability vector of soft cluster assignment.

[5] K. G. Dizaji, A. Herandi, C. Deng, et al. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. ICCV2017.
[6] Junyuan Xie, Ross Girshick, Ali Farhadi. Unsupervised deep embedding for clustering analysis. ICML2016.

OUR UNCOVERING STRATEGY

- **Structural source regularization:** Replacing the auxiliary distribution with that formed by ground-truth labels of source data implements the structural source regularization via a simple strategy of joint network training:

$$\min_{\boldsymbol{\theta},\boldsymbol{\vartheta}} \mathcal{L}^s_{f\circ\varphi} = -\frac{1}{n_s}\sum_{j=1}^{n_s}\sum_{k=1}^{K} \mathrm{I}[k=y^s_j]\log p^s_{j,k}, \quad (7)$$

$$\min_{\boldsymbol{\theta},\{\boldsymbol{\mu}_k\}_{k=1}^K} \mathcal{L}^s_{\varphi} = -\frac{1}{n_s}\sum_{j=1}^{n_s}\sum_{k=1}^{K} \mathrm{I}[k=y^s_j]\log \widetilde{p}^s_{j,k}, \quad (8)$$

where

$$\widetilde{p}^s_{j,k} = \frac{\exp((1+||\boldsymbol{z}^s_j-\boldsymbol{\mu}_k||^2)^{-1})}{\sum_{k'=1}^K \exp((1+||\boldsymbol{z}^s_j-\boldsymbol{\mu}_{k'}||^2)^{-1})}. \quad (9)$$

- We also enhance structural regularization with soft selection of less divergent source examples:

$$w^s(\boldsymbol{x}^s) = \frac{1}{2}\left(1+\frac{\boldsymbol{c}^{t\top}_{y^s}\boldsymbol{x}^s}{||\boldsymbol{c}^t_{y^s}||\,||\boldsymbol{x}^s||}\right)\in[0,1]. \quad (12)$$

where $\{c^t_k\}_{k=1}^K$ are the K target cluster centers in the feature space.

$$\mathcal{L}^s_{f\circ\varphi(\cdot;\{w^s_j\}_{j=1}^{n_s})} = -\frac{1}{n_s}\sum_{j=1}^{n_s} w^s_j \sum_{k=1}^{K} \mathrm{I}[k=y^s_j]\log p^s_{j,k}, \quad (13)$$

$$\mathcal{L}^s_{\varphi(\cdot;\{w^s_j\}_{j=1}^{n_s})} = -\frac{1}{n_s}\sum_{j=1}^{n_s} w^s_j \sum_{k=1}^{K} \mathrm{I}[k=y^s_j]\log \widetilde{p}^s_{j,k}. \quad (14)$$

OUR UNCOVERING STRATEGY

- Training algorithm

Algorithm 1 Training algorithm for SRDC, E denotes the training epoch, I denotes the training iteration, B_t and B_s denote the mini-batches.

Input: unlabeled target samples $\mathcal{T} = \{x_i^t\}_{i=1}^{n_t}$; labeled source samples $\mathcal{S} = \{(x_j^s, y_j^s)\}_{j=1}^{n_s}$

Output: $\theta, \vartheta, \{\mu_k\}_{k=1}^K$

Initialize: $\theta, \vartheta, \{\mu_k\}_{k=1}^K$, $q_{i,k}^t = \tilde{q}_{i,k}^t = \mathrm{I}[k = \hat{y}_i^t]$ for $i \in \{1, 2, \ldots, n_t\}$ and $k \in \{1, 2, \ldots, K\}$, $w_j^s = 1$ for $j \in \{1, 2, \ldots, n_s\}$, $E = 1$

while not converge **do**

    for $I \leftarrow 1, MAX\_ITER$ **do**

        Sample B_t and B_s from $\mathcal{T}$ and $\mathcal{S}$

        if E != 1 **then**

            Compute q_{ik}^t and $\tilde{q}_{ik}^t$ by using (2)

        end if

        Update $\theta, \vartheta, \{\mu_k\}_{k=1}^K$ by minimizing (11) on B_t and B_s

    end for

    Compute $\{c_k^t\}_{k=1}^K$ by standard K-means clustering

    Compute $w_j^s = 1, j \in \{1, 2, \ldots, n_s\}$ by using (12)

    Initialize: $\{\mu_k\}_{k=1}^K$

    E = E + 1

end while

EXPERIMENTS

- Ablation study:

Method	A →W	A →D	D →A	W →A	Avg
Source Model	77.8±0.2	82.1±0.2	64.5±0.2	66.1±0.2	72.6
SRDC (w/o structural source regularization)	87.3±0.0	92.1±0.1	73.9±0.1	75.0±0.1	82.1
SRDC (w/o feature discrimination)	94.2±0.4	94.3±0.4	74.3±0.2	75.5±0.4	84.6
SRDC (w/o soft source sample selection)	94.8±0.2	94.6±0.3	74.6±0.3	75.7±0.3	84.9
SRDC	**95.7**±0.2	**95.8**±0.2	**76.7**±0.3	**77.1**±0.1	**86.3**

Each component is important!

- Convergence performance

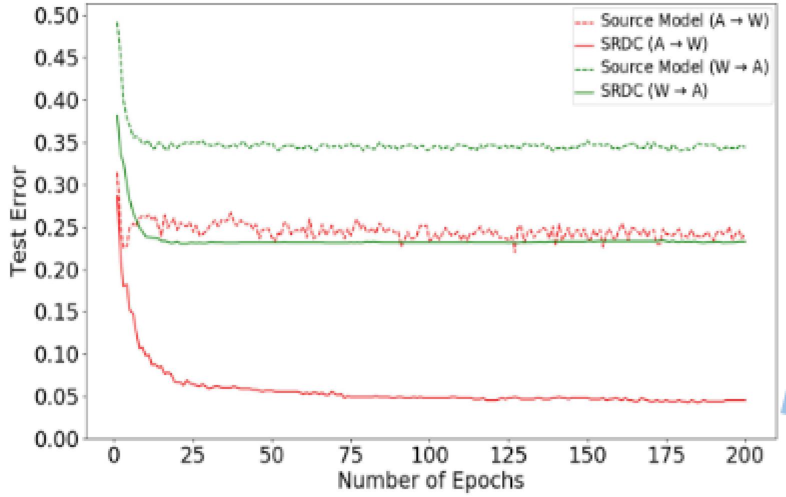


Figure 5. Convergence.

Better and more stable!

- Comparative experiments under inductive UDA setting:

Method	A →W	A →D	D →A	W →A	Avg
Source Model	79.3	81.6	63.1	65.7	72.4
DANN [16]	80.8	82.4	66.0	64.6	73.5
MCD [48]	86.5	86.7	72.4	70.9	79.1
SRDC	**91.9**	**91.6**	**75.6**	**75.7**	**83.7**
Oracle Model	98.8	97.6	87.8	87.8	93.0

Much closer to Oracle Model! Stronger generalization ability!

EXPERIMENTS

- Source refinement

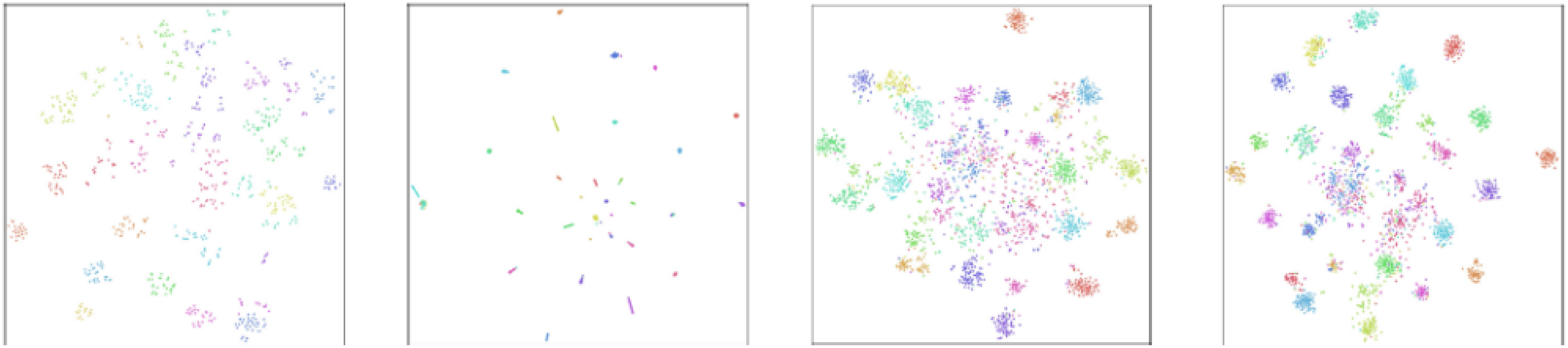
From canonical viewpoint

From canonical, top-down, bottom-up, and side viewpoints

Figure 2. The images on the left are randomly sampled from the target domain **A** and those on the right are the top-ranked (the 3^{rd} column) and bottom-ranked (the 4^{th} column) samples from the source domain **W** for three classes. Note that the red numbers are the source weights computed by (12).

EXPERIMENTS

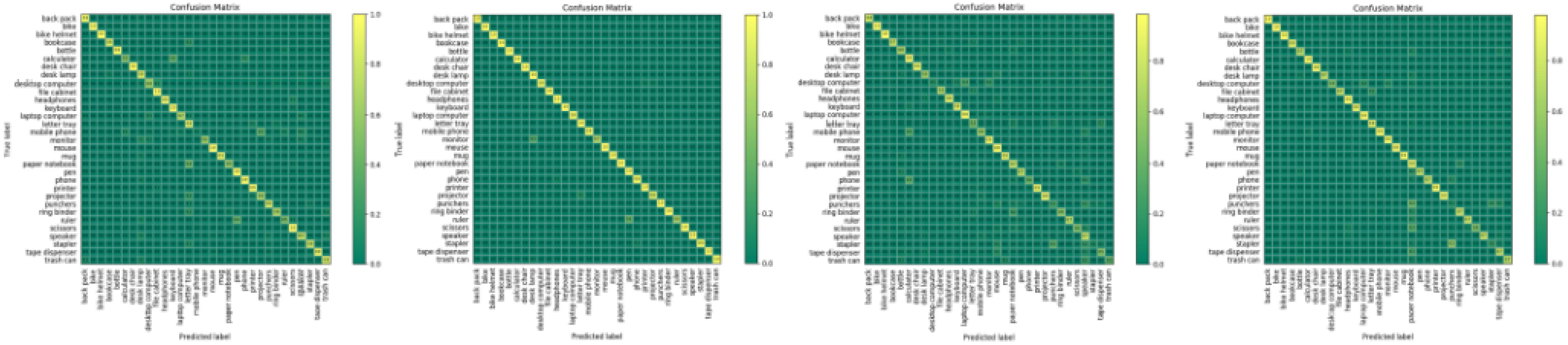
- Visualization by t-SNE and confusion matrix



(a) Source Model: **A→W** (b) SRDC: **A→W** (c) Source Model: **W→A** (d) SRDC: **W→A**

Figure 3. The t-SNE visualization of embedded features on the target domain. Note that different classes are denoted by different colors.

Significant improvement!



(a) Source Model: **A→W** (b) SRDC: **A→W** (c) Source Model: **W→A** (d) SRDC: **W→A**

Figure 4. The confusion matrix on the target domain. (Zoom in to see the exact class names!)

EXPERIMENTS

- Comparison with SOTA

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
Source Model [21]	77.8±0.2	96.9±0.1	99.3±0.1	82.1±0.2	64.5±0.2	66.1±0.2	81.1
MDD [66]	94.5±0.3	98.4±0.1	**100.0**±0.0	93.5±0.2	74.6±0.3	72.2±0.1	88.9
CAN [27]	94.5±0.3	99.1±0.2	99.8±0.2	95.0±0.3	**78.0**±0.3	77.0±0.3	90.6
SRDC	**95.7**±0.2	99.2±0.1	**100.0**±0.0	**95.8**±0.2	76.7±0.3	**77.1**±0.1	**90.8**

Table 3. Results (%) on Office-31 (ResNet-50).

Outperform SOTA!

Methods	I → P	P → I	I → C	C → I	C → P	P → C	Avg
Source Model [21]	74.8±0.3	83.9±0.1	91.5±0.3	78.0±0.2	65.5±0.3	91.2±0.3	80.7
SAFN+ENT [60]	79.3±0.1	93.3±0.4	96.3±0.4	91.7±0.0	77.6±0.1	95.3±0.1	88.9
SymNets [68]	80.2±0.3	93.6±0.2	97.0±0.3	93.4±0.3	78.7±0.3	96.4±0.1	89.9
SRDC	**80.8**±0.3	**94.7**±0.2	**97.8**±0.2	**94.1**±0.2	**80.0**±0.3	**97.7**±0.1	**90.9**

Table 4. Results (%) on ImageCLEF-DA (ResNet-50).

Methods	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
Source Model [21]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
SymNets [68]	47.7	72.9	78.5	64.2	71.3	74.2	64.2	48.8	79.5	74.5	52.6	82.7	67.6
MDD [66]	**54.9**	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	**60.2**	82.3	68.1
SRDC	52.3	**76.3**	**81.0**	**69.5**	**76.2**	**78.0**	**68.7**	**53.8**	**81.7**	**76.3**	57.1	**85.0**	**71.3**

Table 5. Results (%) on Office-Home (ResNet-50).

Code

FUTURE DIRECTIONS

- To investigate more effective clustering methods for target discrimination.
- To discover more helpful manners to enforce structural source regularization.
- To explore novel UDA paradigms with theoretical guidance, which can preserve the intrinsic target discrimination as much as possible.

Thanks for listening !