IEEE International Conference on Multimedia and Expo. 2016

Recognizing Heterogeneous Cross-Domain Data via Generalized Joint Distribution Adaptation

Yuan-Ting Hsieh[1], Shi-Yen Tao[1],

Yao-Hung Hubert Tsai[2], Yi-Ren Yeh[3], and Yu-Chiang Frank Wang[2]

[1]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan

[2]Research Center for IT Innovation, Academia Sinica, Taipei, Taiwan

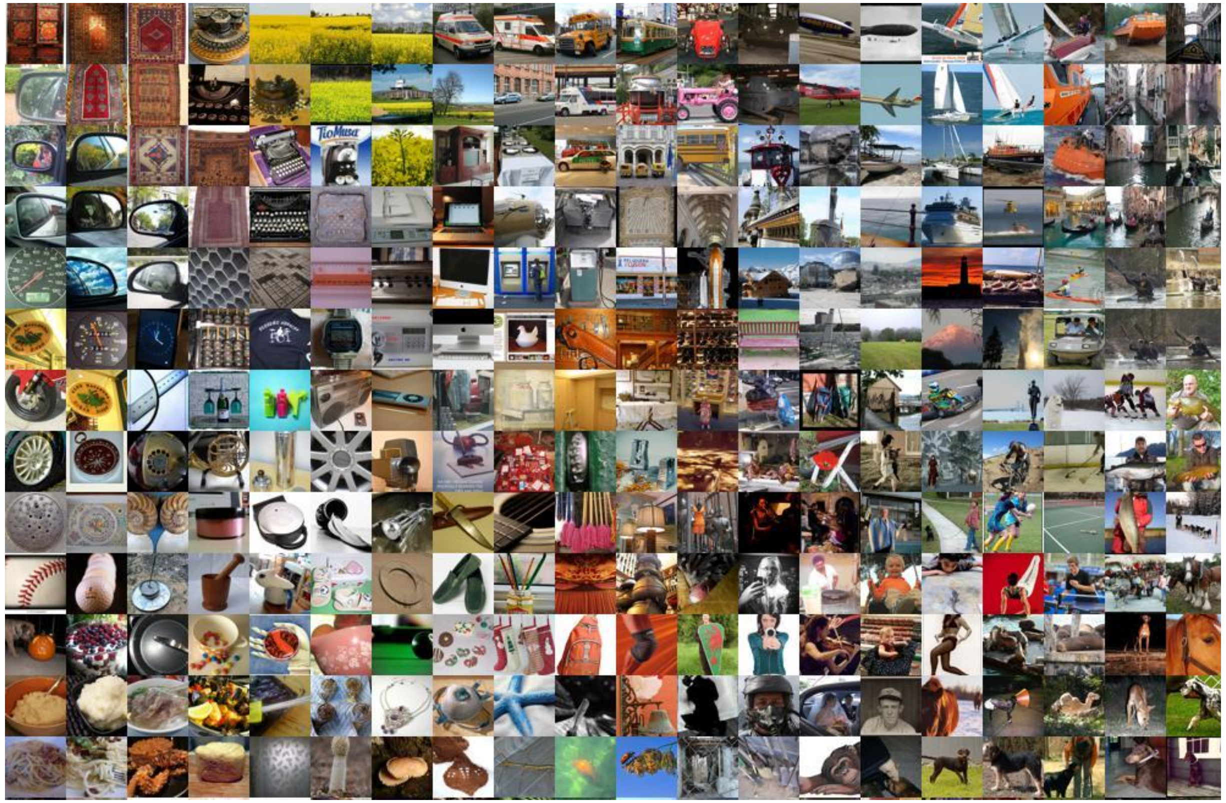
[3]Department of Mathematics, National Kaohsiung Normal University, Kaohsiung, Taiwan

Outline

- Motivation
- Domain Adaptation
- Proposed Method
- Experiment
- Conclusion

Motivation

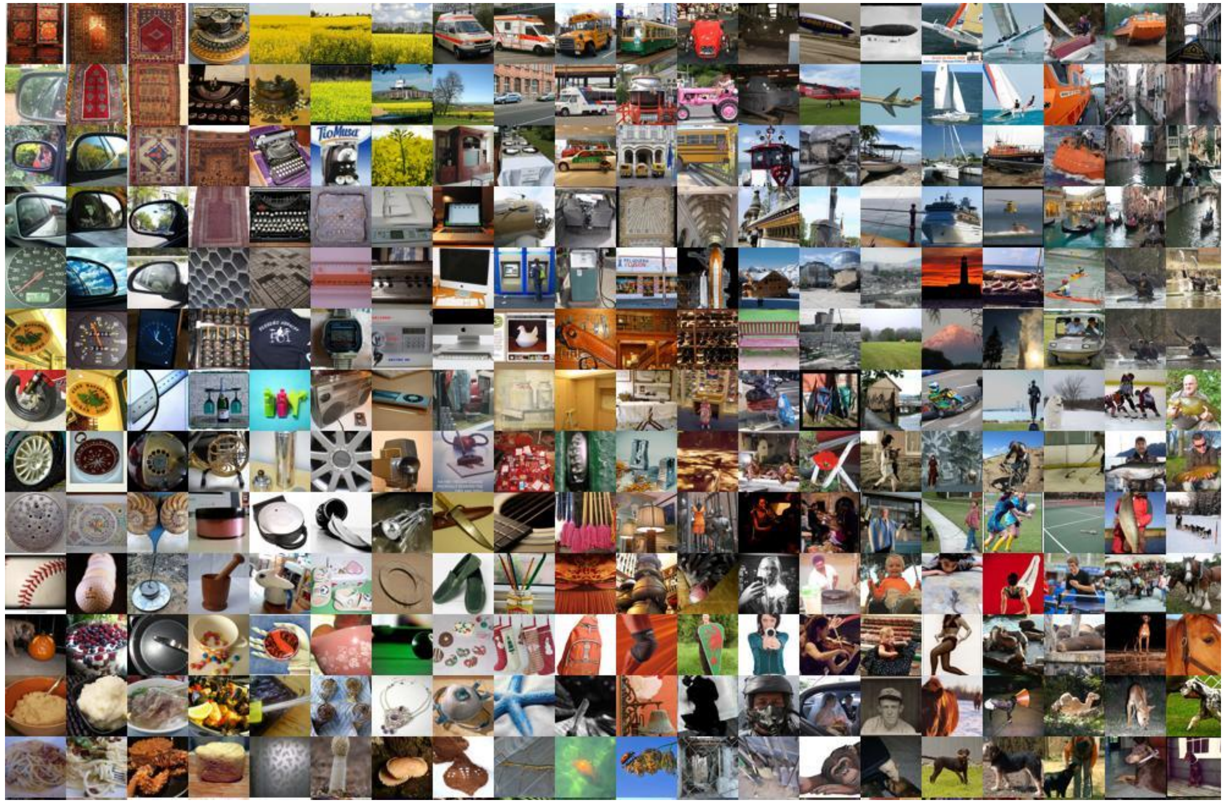
- Labeling a huge amount of data for training is time-consuming!



*Courtesy Andrej Karpathy

Motivation (cont'd)

- Labeling a huge amount of data for training is time-consuming!
  - Can take data from one dataset, and apply the learned classifiers for another??



*Courtesy Andrej Karpathy

Motivation (cont'd)

- Labeling a huge amount of data for training is time-consuming!
    - Can take data from one dataset, and apply the learned classifiers for another?? ***Not A Good Idea!!***
    - Possible ***mismatch*** between datasets/domains (see example below)

Source Domain/Dataset

Target Domain/Dataset

Examples of Domain Mismatch

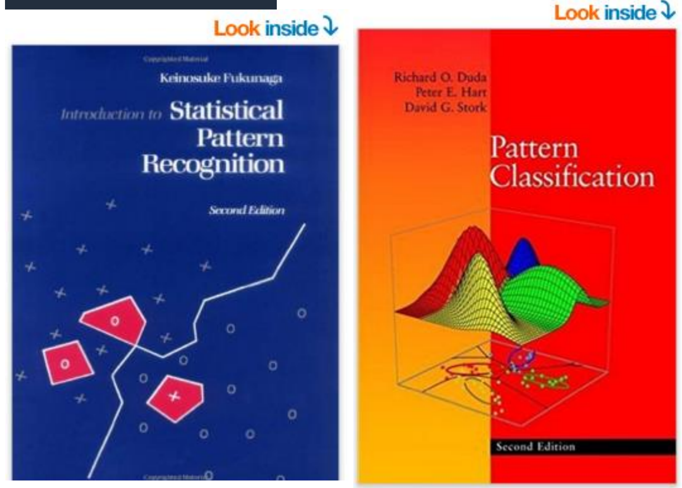
Music Retrieval

Source → **Target**

Examples of Domain Mismatch (cont'd)

Semantic Analysis

amazon Try Prime



★★★★☆ **pattern recognition in engineering**

By Michael R. Chernick on February 8, 2008

Format: Hardcover

Fukunaga is a standard source for pattern recognition methods often cited in the engineering literature. Covers parametric (particularly linear and quadratic discriminant algorithms) and nonparametric methods (density estimation). It is designed for and popular with engineers. When I was working at Nichols Research Corporation Fukunaga's papers and this book (earlier edition) were often cited as sources to justify the algorithms we used for discrimination problems. In fact Fukunaga had been a consultant to the company (used primarily by the Boston branch of the company where the KENN algorithms were developed). It is a reputable source. I still like Duda and Hart (1973) for good explanations of the fundamental concepts. The second edition that was recent ly published with Stark as a third author is also highly recommended. For statisticians McLachlan's book is now far and away the best source.

Comment | 31 people found this helpful. Was this review helpful to you? Yes No Report abuse

Rotten Tomatoes

While perfectly serviceable as an hour-and-a-half of shocks and scares, there's substance missing to The Boy that prevents it from truly coming to life.

Full Review... | March 17, 2016

Source → **Target**

What Can We Do?

- To solve the above cross-domain recognition tasks…
  - **Domain Adaptation**
  - Aim to address the (same) learning task across different domains

Source Domain/Dataset

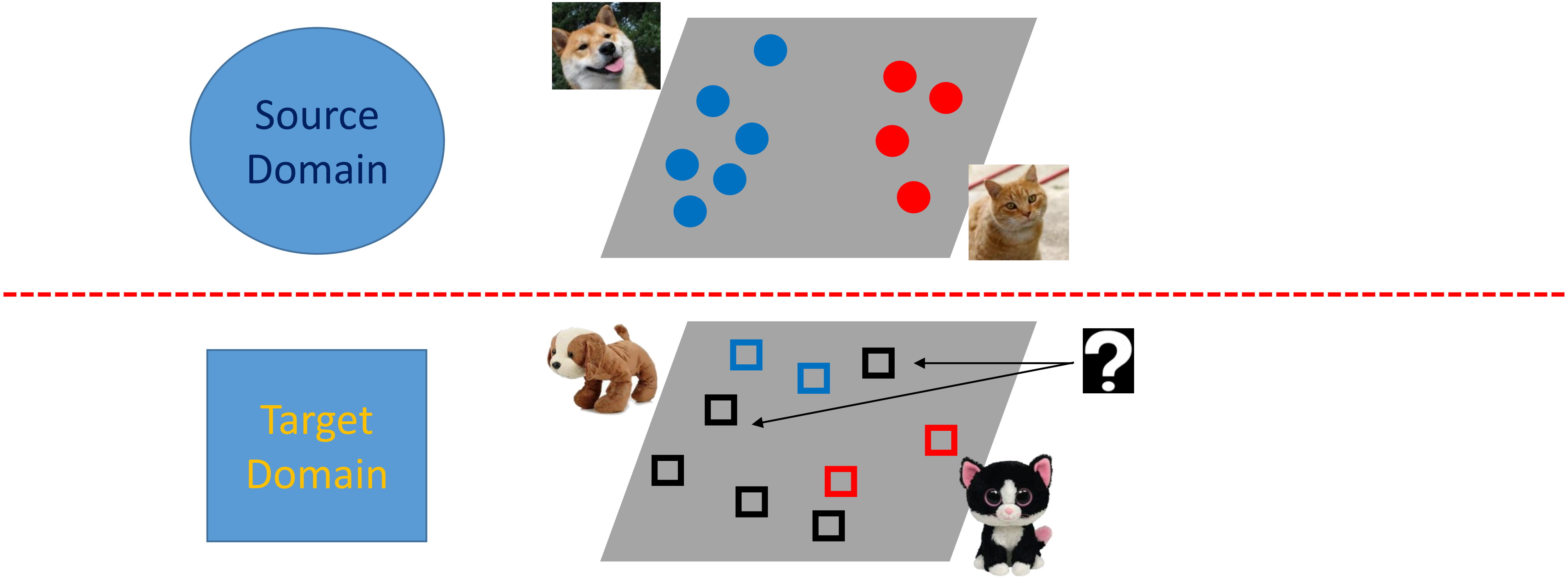
Target Domain/Dataset

Outline

- Motivation
- Domain Adaptation (DA)
  - Semi-Supervised vs. Unsupervised DA
- Proposed Method
- Experiment
- Conclusion

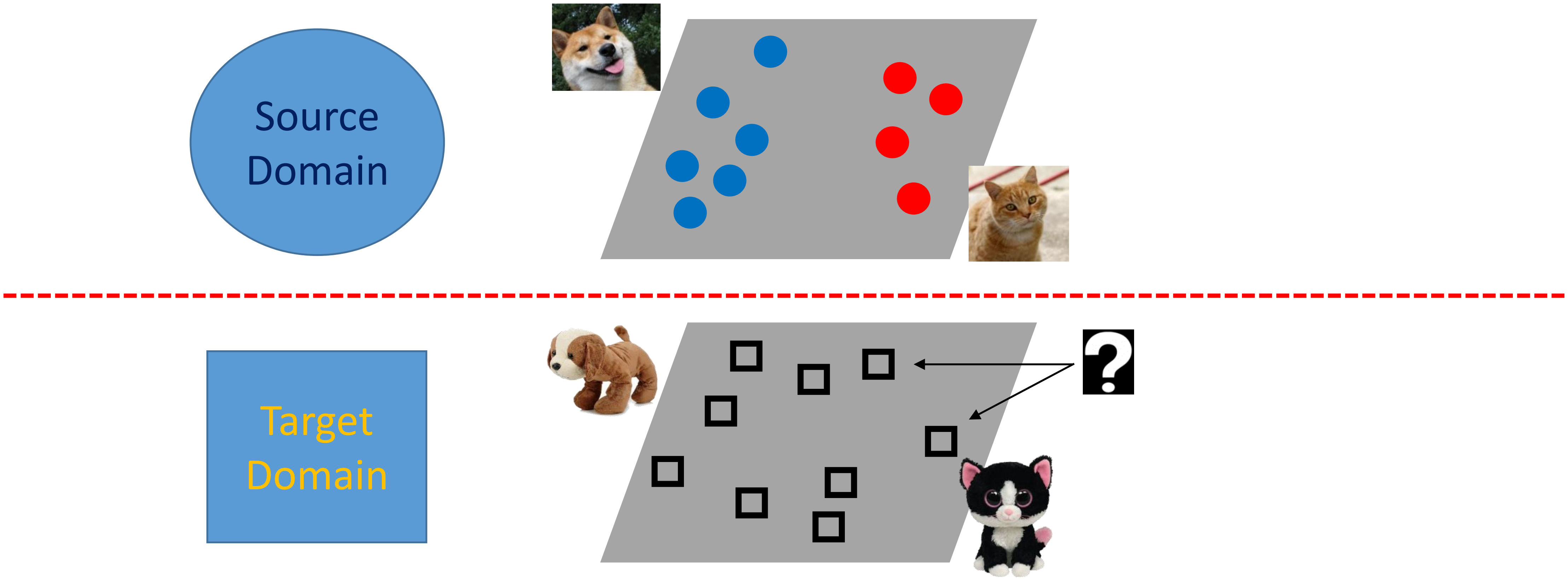
Semi-Supervised DA

- Source domain: label-data available
- Target domain: few label data, most are unlabeled



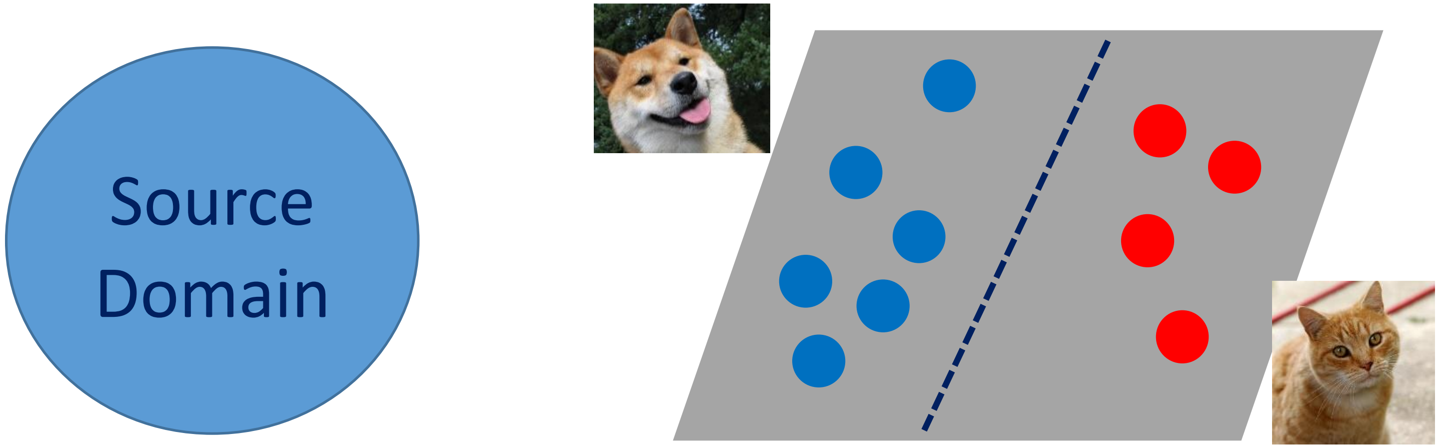
Unsupervised DA

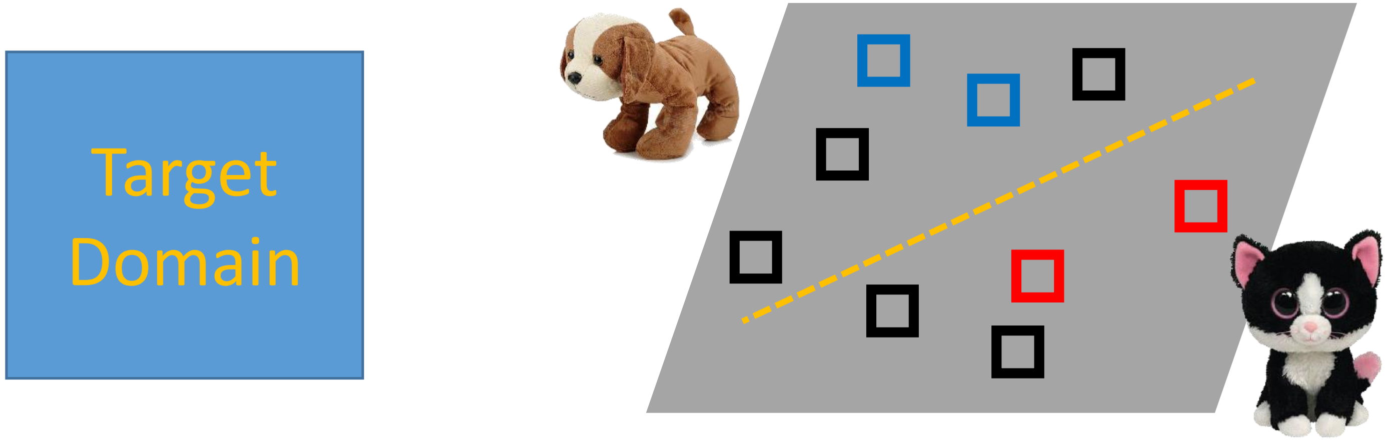
- Source domain: label-data available
- Target domain: only unlabeled data are presented



Goal of Semi-Supervised/Unsupervised DA

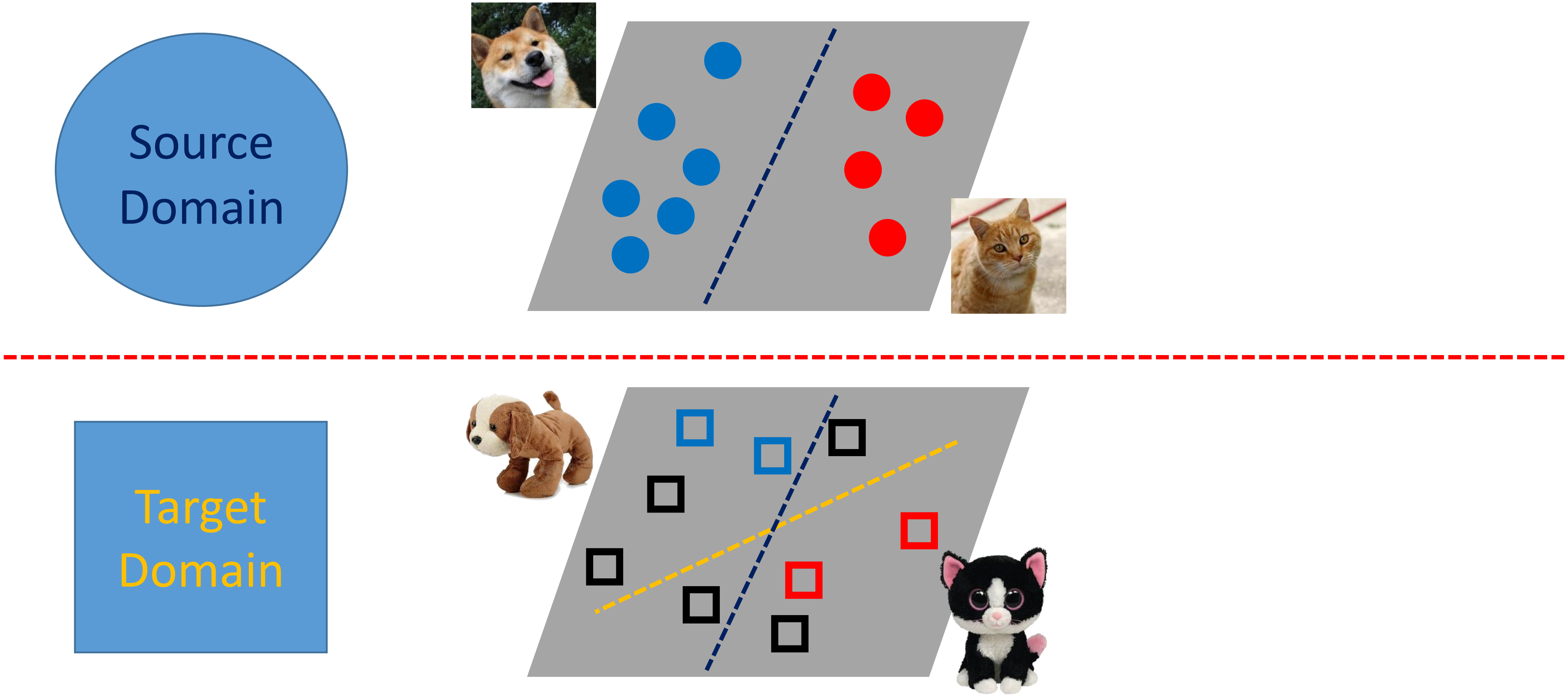
- Adapt the learned model from source to target domains





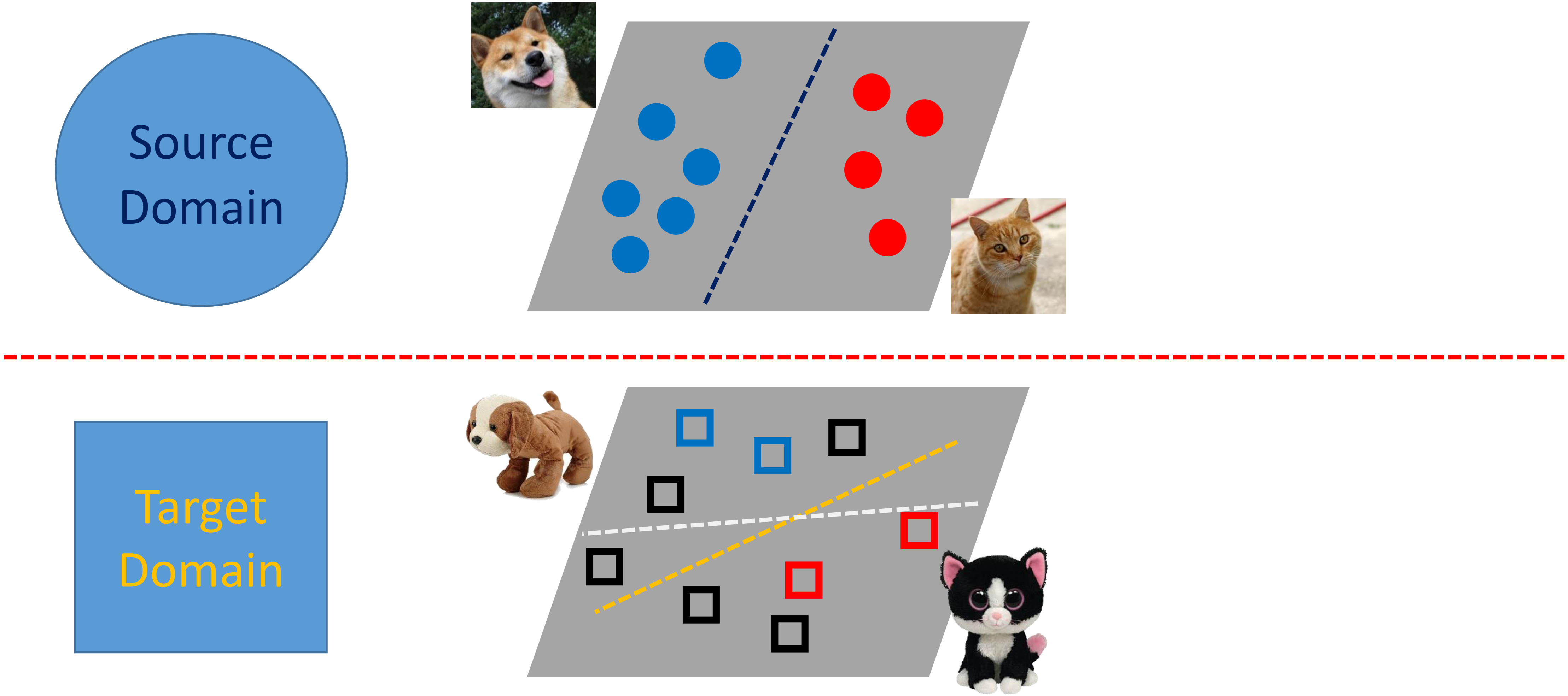
Goal of Semi-Supervised/Unsupervised DA (cont’d)

- Adapt the learned model from source to target domains



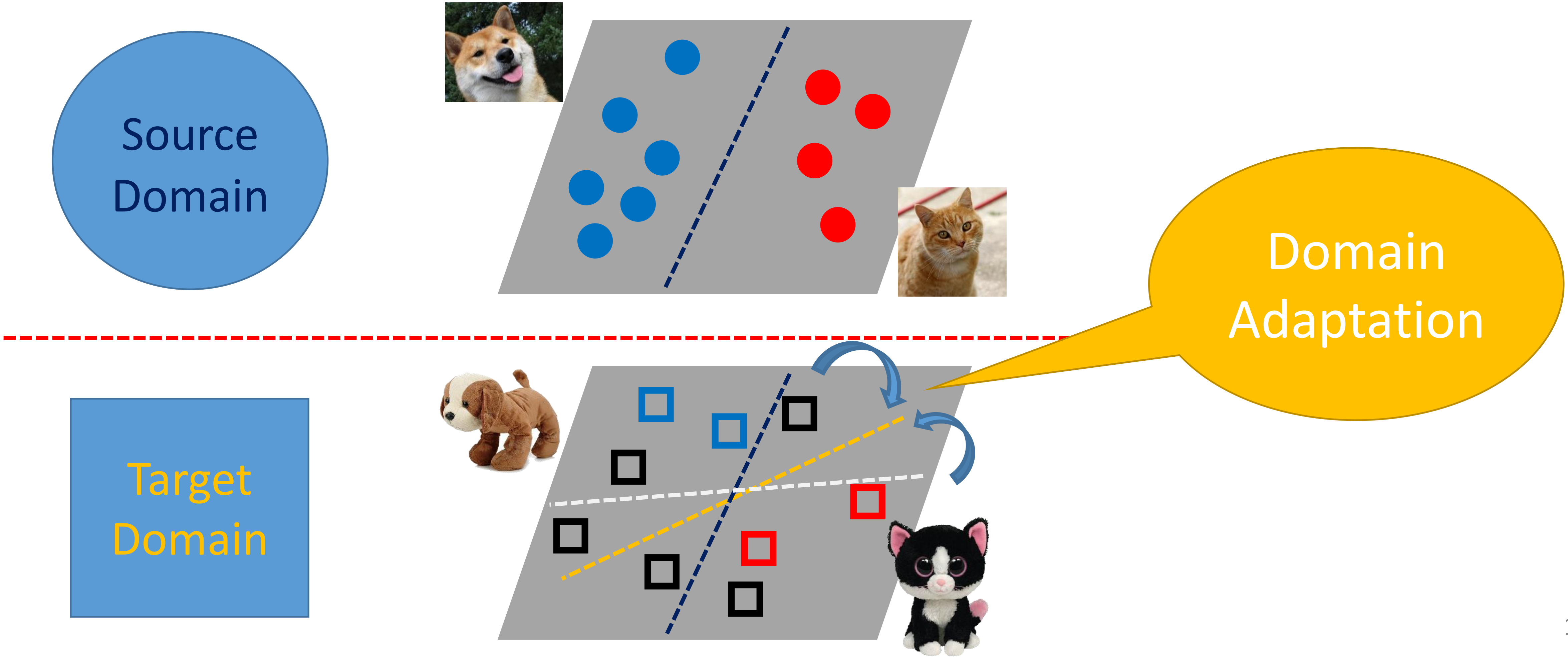
Goal of Semi-Supervised/Unsupervised DA (cont'd)

- Adapt the learned model from source to target domains



Goal of Semi-Supervised/Unsupervised DA (cont’d)

- Adapt the learned model from source to target domains

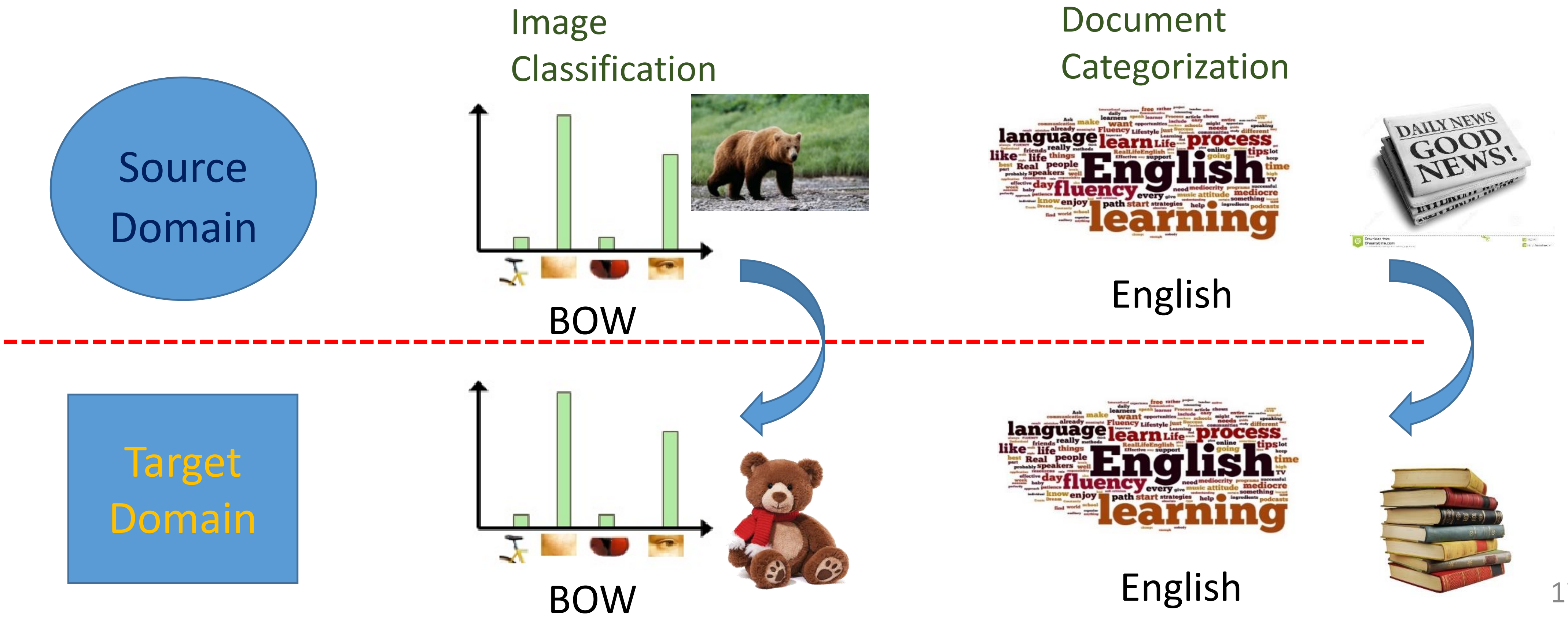


Outline

- Motivation
- Domain Adaptation
  - Semi-Supervised vs. Unsupervised DA
  - Homogeneous vs. Heterogeneous DA
- Proposed Method
- Experiment
- Conclusion

Homogeneous DA

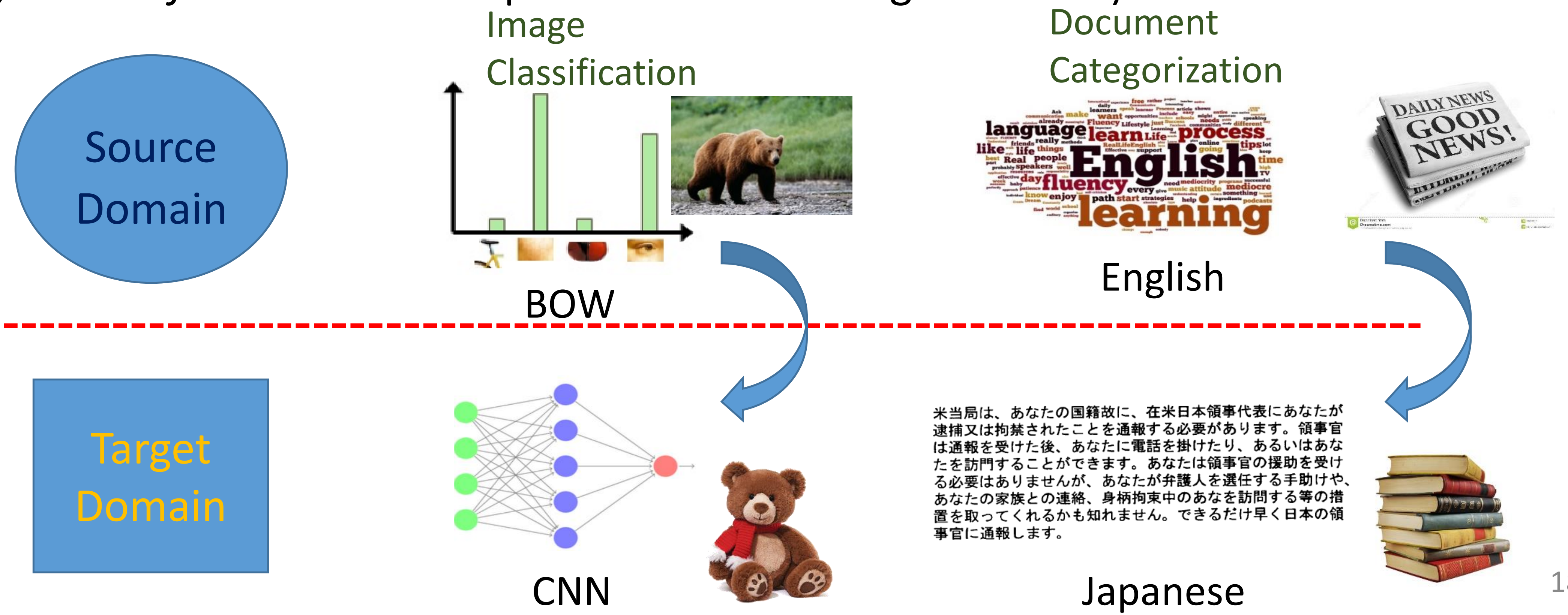
- Source and target-domain data are with the same feature type
- Can be applied to semi-supervised or even unsupervised DA



Heterogeneous DA

- Source and target-domain data are with the different feature types/dims
- Typically require semi-supervised DA settings

  (i.e., at least *few* labeled data presented in the target domain)



What we address in our work…

- Semi-supervised & Heterogeneous DA
- With applications to

1. Cross-Domain Object Recognition
2. Cross-Lingual Text Categorization

米当局は、あなたの国籍故に、在米日本領事代表にあなたが
逮捕又は拘禁されたことを通報する必要があります。領事官
は通報を受けた後、あなたに電話を掛けたり、あるいはあな
たを訪門することができます。あなたは領事官の援助を受け
る必要はありませんが、あなたが弁護人を選任する手助けや、
あなたの家族との連絡、身柄拘束中のあなを訪問する等の措
置を取ってくれるかも知れません。できるだけ早く日本の領
事官に通報します。

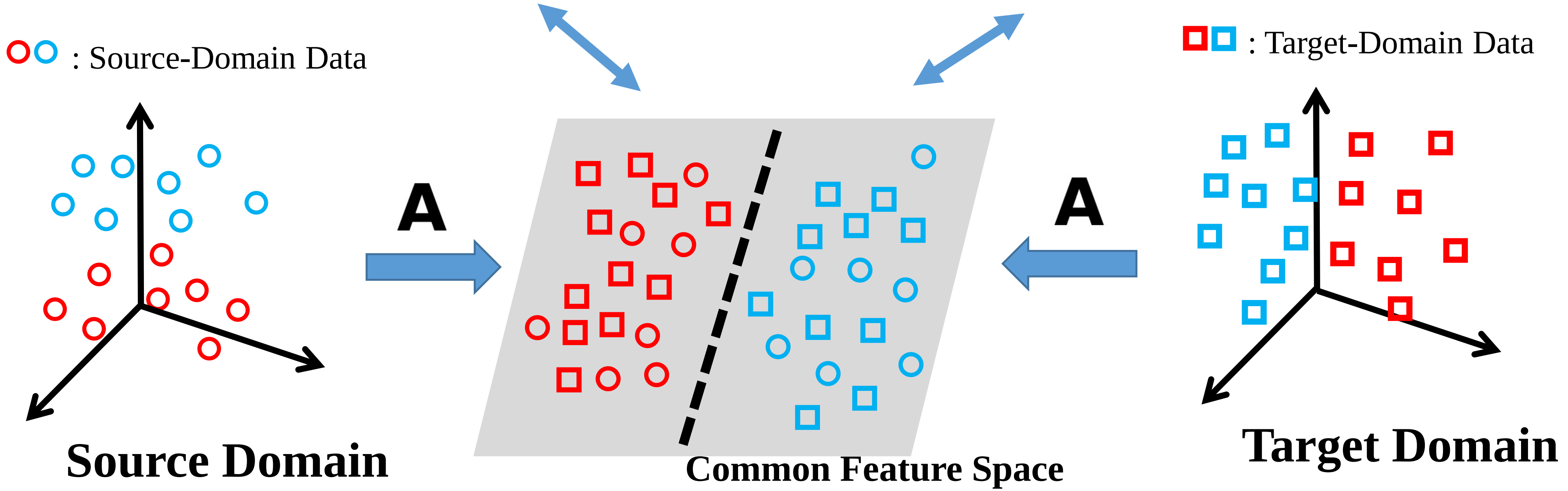
Outline

- Motivation
- Domain Adaptation
- Proposed Method
- Experiment
- Conclusion

Motivation: Joint Distribution Adaptation

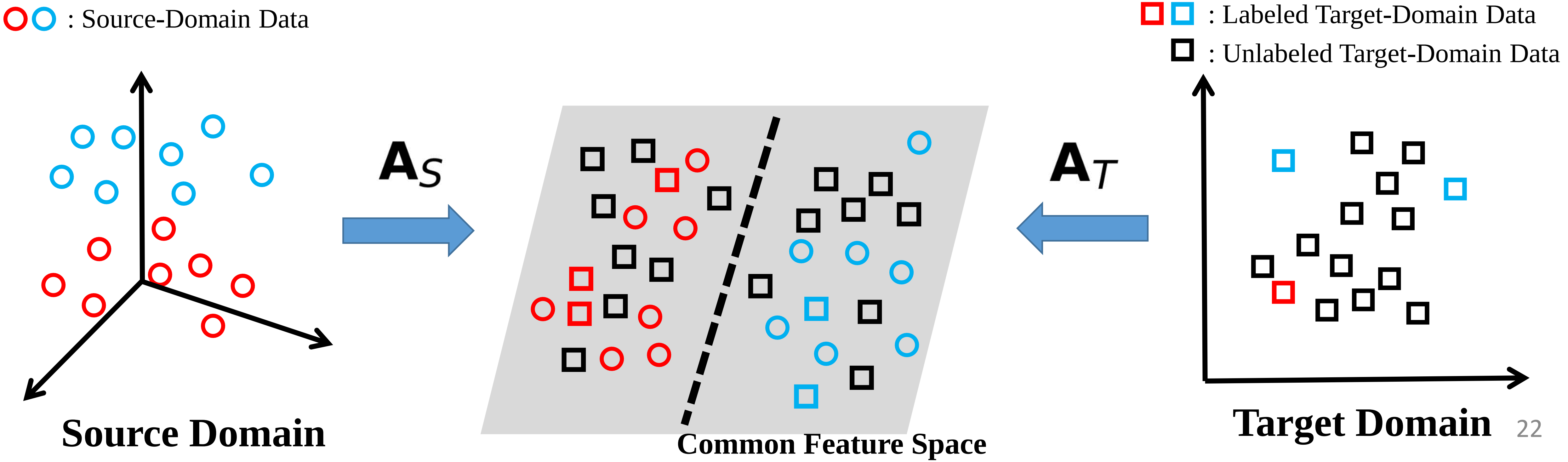
- Joint Distribution Adaptation (JDA) for homogeneous DA [Long *et al.*, ICCV'13]
- Find a proper common feature space via transformation matrix $\mathbf{A}$, so that both marginal $\mathcal{P}(\mathbf{X})$ and conditional distributions $\mathcal{P}(\mathbf{X}|\mathbf{Y})$ can be matched.

$$\mathcal{P}_S(\mathbf{A}^\top \mathbf{X}_S) \approx \mathcal{P}_T(\mathbf{A}^\top \mathbf{X}_T) \qquad \mathcal{P}_S(\mathbf{A}^\top \mathbf{X}_S | Y_S) \approx \mathcal{P}_T(\mathbf{A}^\top \mathbf{X}_T | Y_T)$$



Our Proposed Method

- **Generalized Joint Distribution Adaptation (G-JDA)**
  - JDA only addresses homogeneous DA & does not utilize any target-domain labels
- For semi-supervised & heterogeneous DA
  - Derive transforms $\mathbf{A}_S$ and $\mathbf{A}_T$ for relating source and target domain data.



Notation

- Source domain data: $\mathcal{D}_S = \{\mathbf{X}_S, Y_S\} = \{\mathbf{x}_S^i, y_S^i\}_{i=1}^{n_S}$, $\mathbf{x}_S^i \in \mathbb{R}^{d_S}$
- Target domain labelled data: $\mathcal{D}_L = \{\mathbf{X}_L, Y_L\} = \{\mathbf{x}_L^i, y_L^i\}_{i=1}^{n_L}$, $\mathbf{x}_L^i \in \mathbb{R}^{d_T}$
- Target domain unlabeled data: $\mathcal{D}_U = \{\mathbf{X}_U, Y_U\} = \{\mathbf{x}_U^i, y_U^i\}_{i=1}^{n_U}$, $\mathbf{x}_U^i \in \mathbb{R}^{d_T}$
- Note that $\mathbb{R}^{d_S} \neq \mathbb{R}^{d_T}$
- Assume that both domains have the same C classes of interest.

Distribution Alignment for G-JDA

- G-JDA aims to find a common feature space where both marginal & conditional distributions are matched.

$$\mathcal{P}_S(\mathbf{A}_S^\top \mathbf{X}_S) \approx \mathcal{P}_T(\mathbf{A}_T^\top \mathbf{X}_T) \qquad \mathcal{P}_S(\mathbf{A}_S^\top \mathbf{X}_S | Y_S) \approx \mathcal{P}_T(\mathbf{A}_T^\top \mathbf{X}_T | Y_T)$$

- Following JDA, we apply MMD to measure distribution discrepancy. [A. Gretton *et al*., NIPS 2006]

Marginal distribution discrepancy $\Rightarrow$

$$E_{mar}(\mathbf{A}_S, \mathbf{A}_T) = \| \frac{1}{n_S} \sum_{i=1}^{n_S} \mathbf{A}_S^\top \mathbf{x}_S^i - \frac{1}{n_T} \sum_{j=1}^{n_T} \mathbf{A}_T^\top \mathbf{x}_T^j \|^2$$

Conditional distribution discrepancy $\Rightarrow$

$$E_{cond}^{(c)}(\mathbf{A}_S, \mathbf{A}_T) = \| \frac{1}{n_S^c} \sum_{i=1}^{n_S^c} \mathbf{A}_S^\top \mathbf{x}_S^{i,c} - \frac{1}{n_T^c} \sum_{j=1}^{n_T^c} \mathbf{A}_T^\top \mathbf{x}_T^{j,c} \|^2$$

Objective Function of G-JDA

- Derive transforms $\mathbf{A}_S$ and $\mathbf{A}_T$ for relating source & target domain data.
- The resulting feature space would better match the marginal and conditional distributions of projected cross-domain data.

$$\min_{\mathbf{A}_S,\mathbf{A}_T} \quad E_{mar}(\mathbf{A}_S,\mathbf{A}_T)+\sum_{c=1}^{C} E_{cond}^{(c)}(\mathbf{A}_S,\mathbf{A}_T)+\lambda\left(\|\mathbf{A}_S\|^2+\|\mathbf{A}_T\|^2\right)$$

$$\text{s.t.} \quad \hat{\mathbf{X}}\mathbf{H}\hat{\mathbf{X}}^\top=\mathbf{I},$$

$\Rightarrow$ Prevent trivial solution

$$\text{where} \quad \hat{\mathbf{X}}=[\mathbf{A}_S^\top\mathbf{X}_S,\mathbf{A}_T^\top\mathbf{X}_T],\ \mathbf{H}=\mathbf{I}_{n_S+n_T}-\frac{1}{n_S+n_T}\mathbf{1}_{n_S+n_T}$$

G-JDA Algorithm

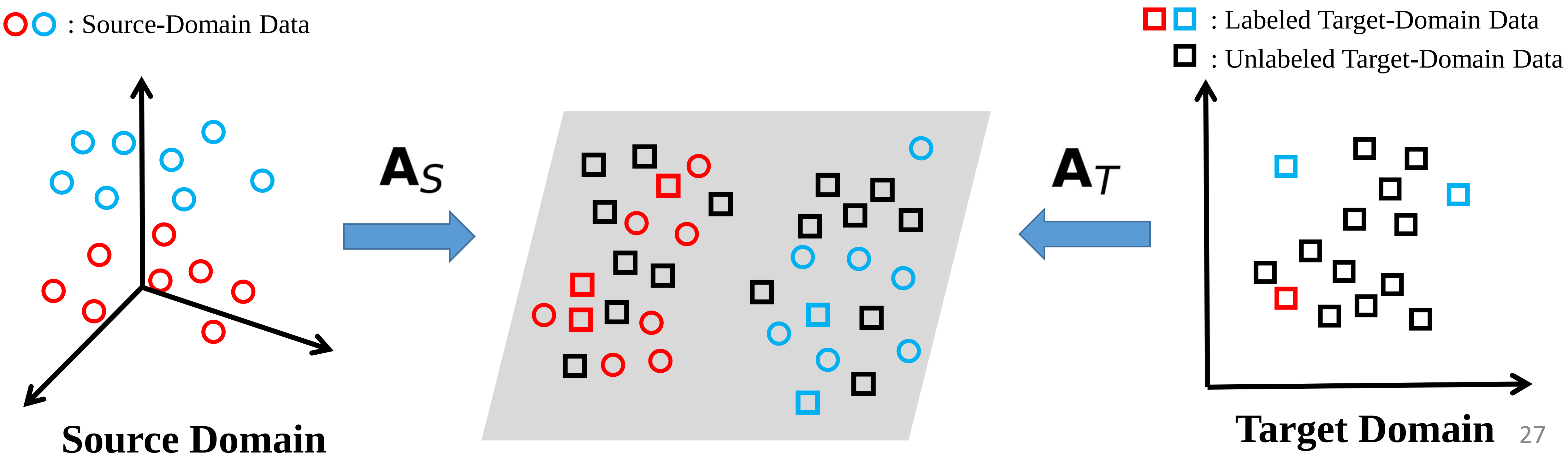
1. Use labeled data in both domains to initialize A_S and A_T.

Source Domain

Target Domain

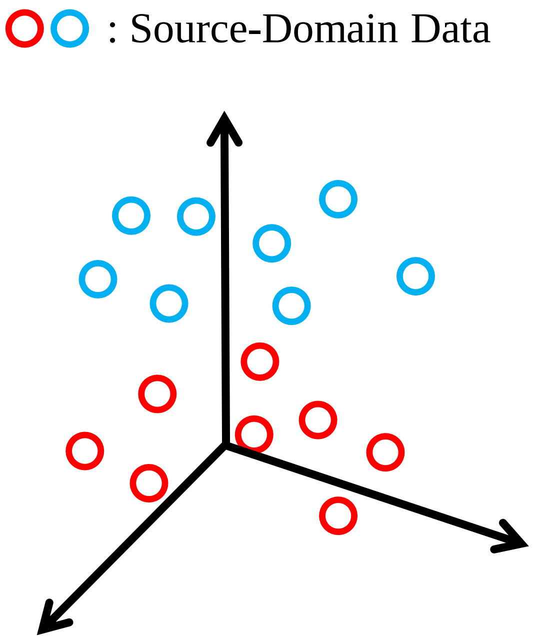
G-JDA Algorithm (cont'd)

2. Project all labeled and unlabeled data onto the resulting feature space.

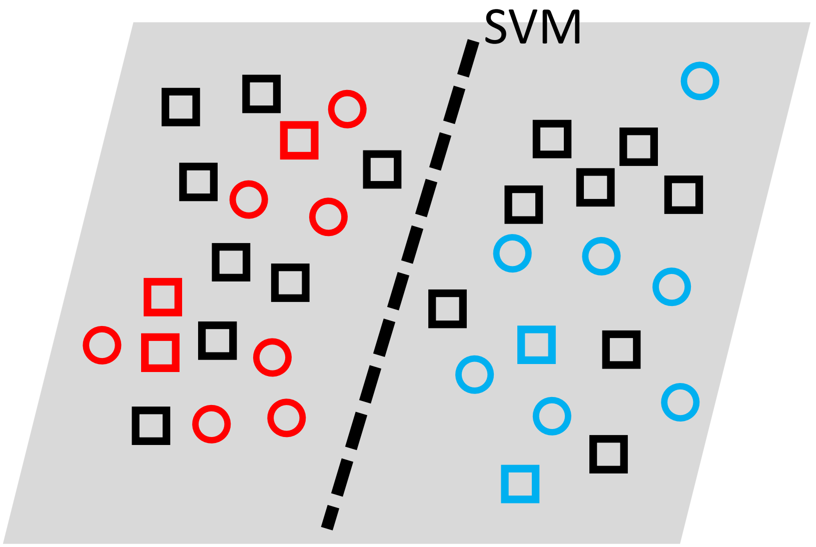


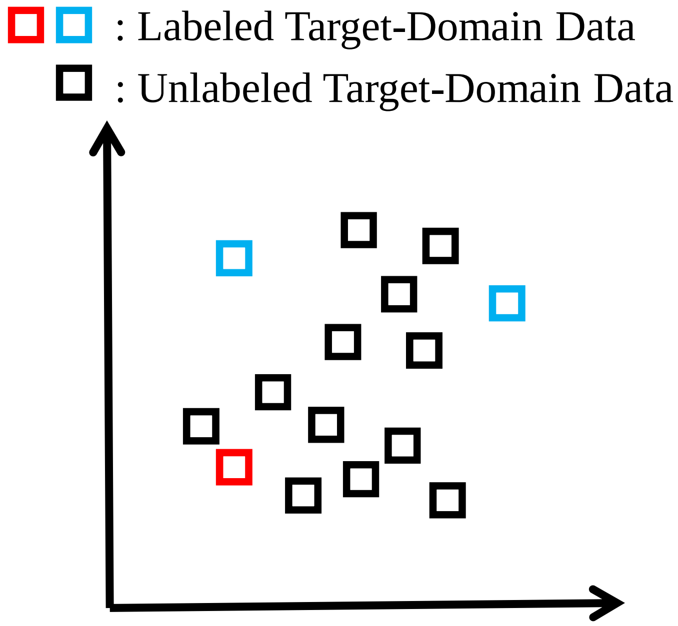
G-JDA Algorithm (cont'd)

3. Train a (linear) SVM in the common feature space.



Source Domain

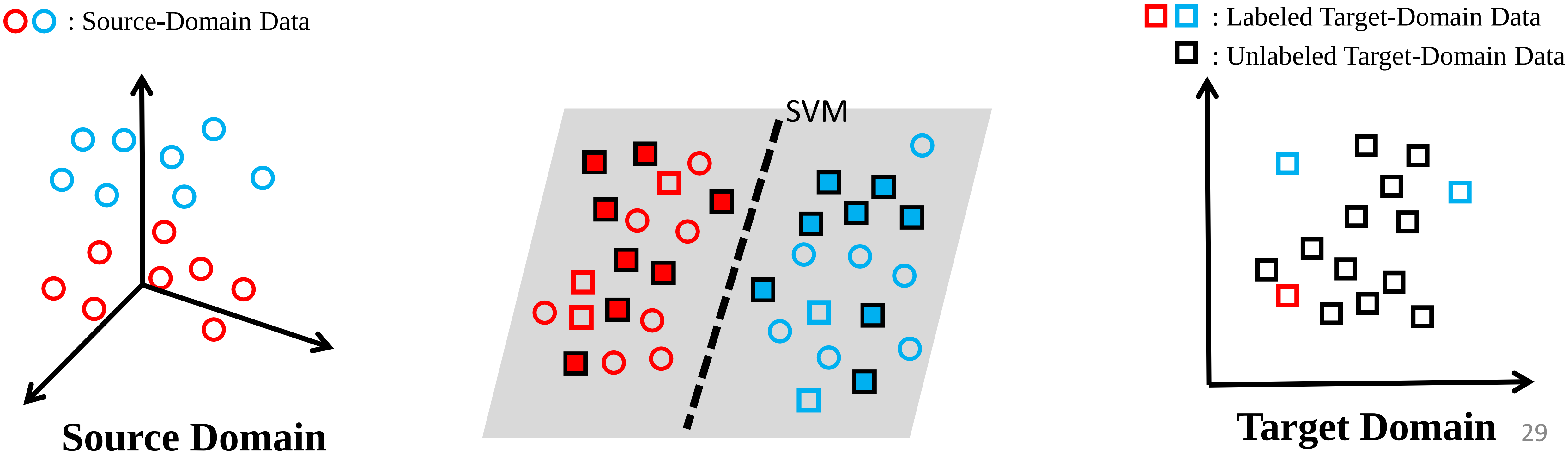




Target Domain

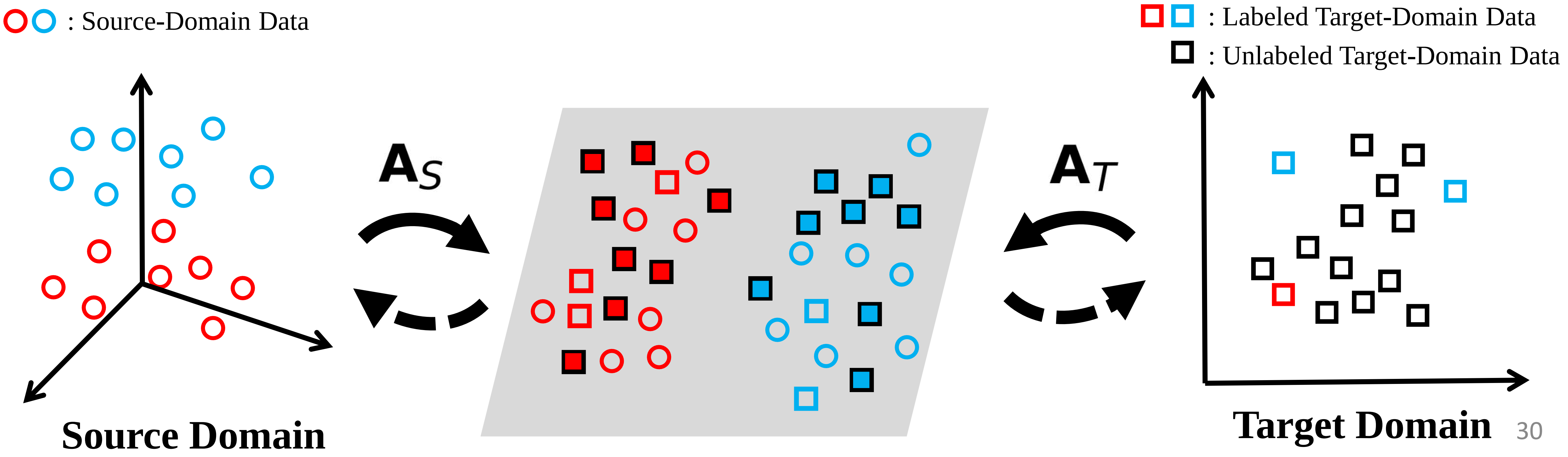
G-JDA Algorithm (cont’d)

4. Predict the pseudo labels for the projected unlabeled target-domain data.



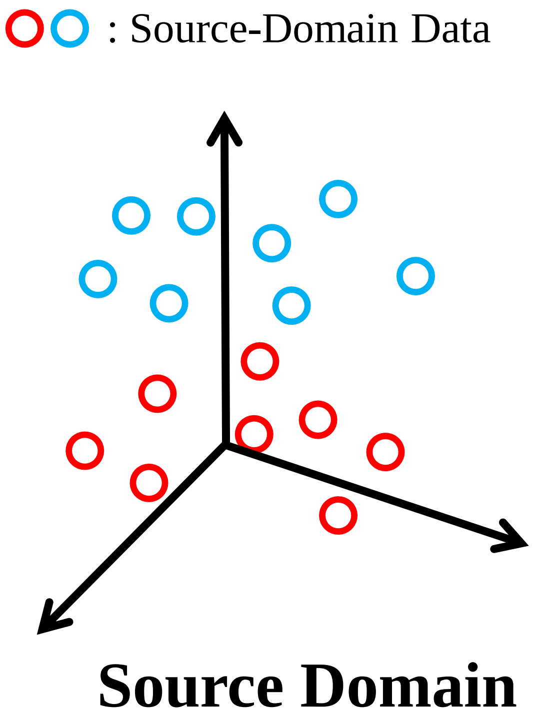
G-JDA Algorithm (cont'd)

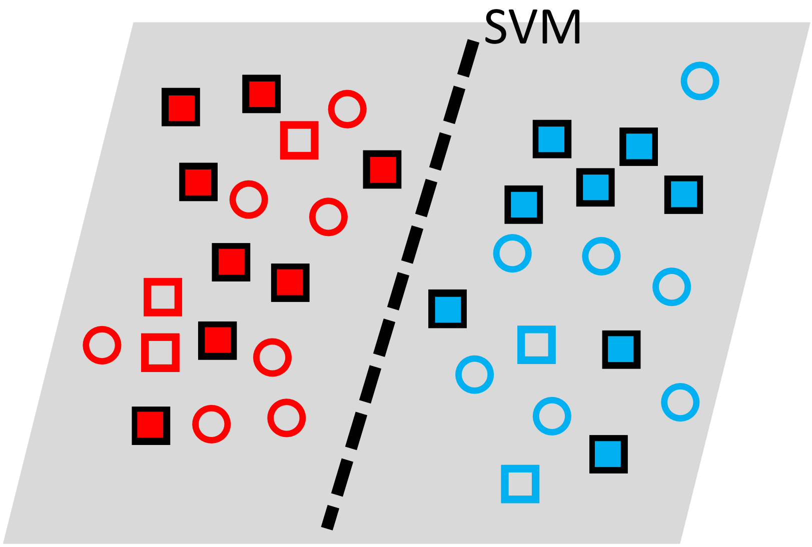
5. Apply exiting and pseudo labels for cross-domain data to update $\mathbf{A}_S$ and $\mathbf{A}_T$.

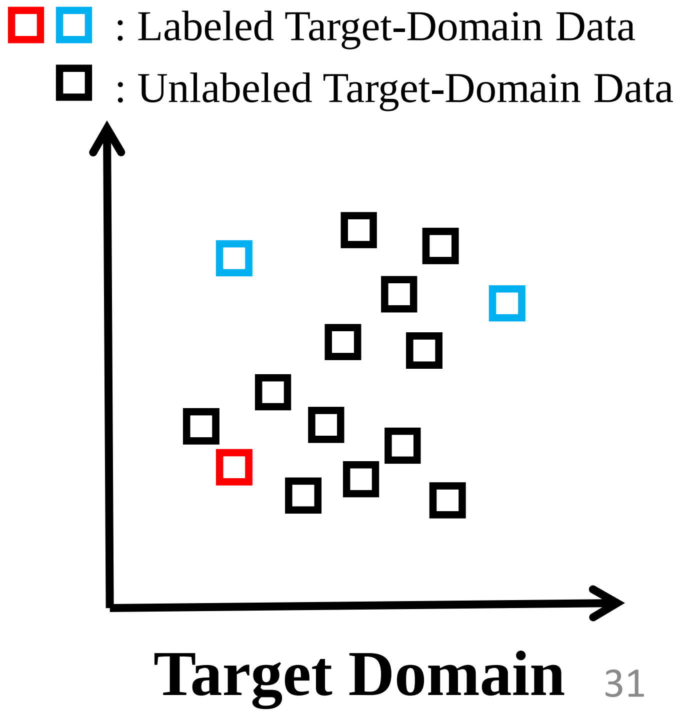


G-JDA Algorithm (cont'd)

6. Update the SVM and the pseudo labels.

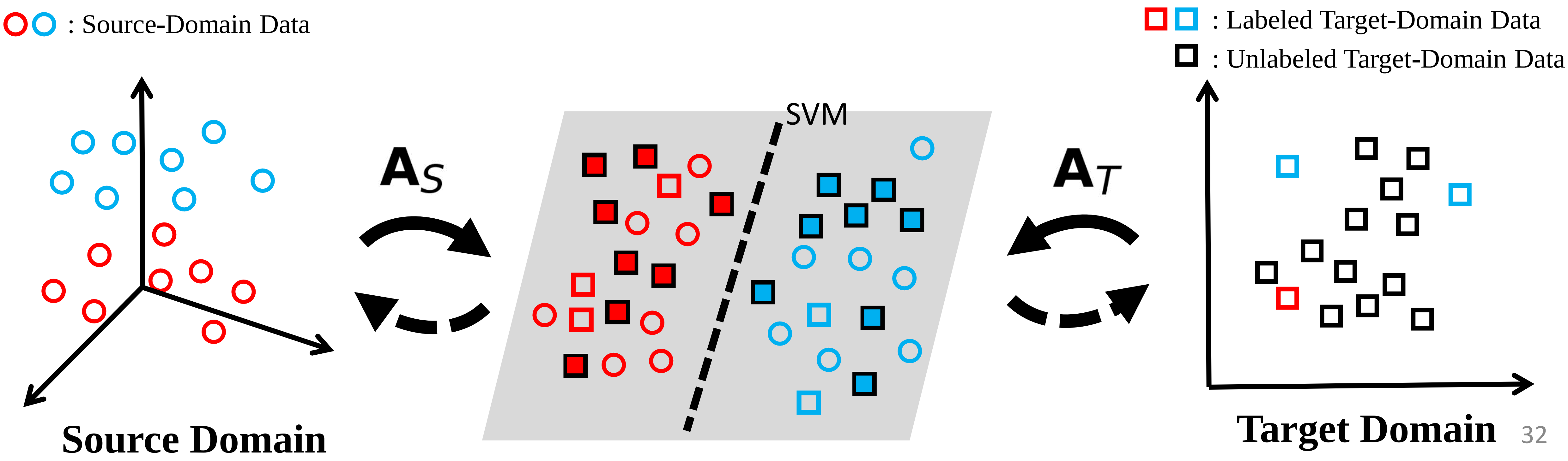






G-JDA Algorithm (cont’d)

7. Repeat the same process until converge.



Outline

- Motivation
- Domain Adaptation
- Proposed Method
- Experiment
- Conclusion

Image Classification Task

- Office and Caltech-256 datasets
- Office: Amazon (A), Webcam (W), DSLR (D) [K. Saenko *et al*., 2010]
- Caltech (C) [G. Griffin *et al*., 2007]

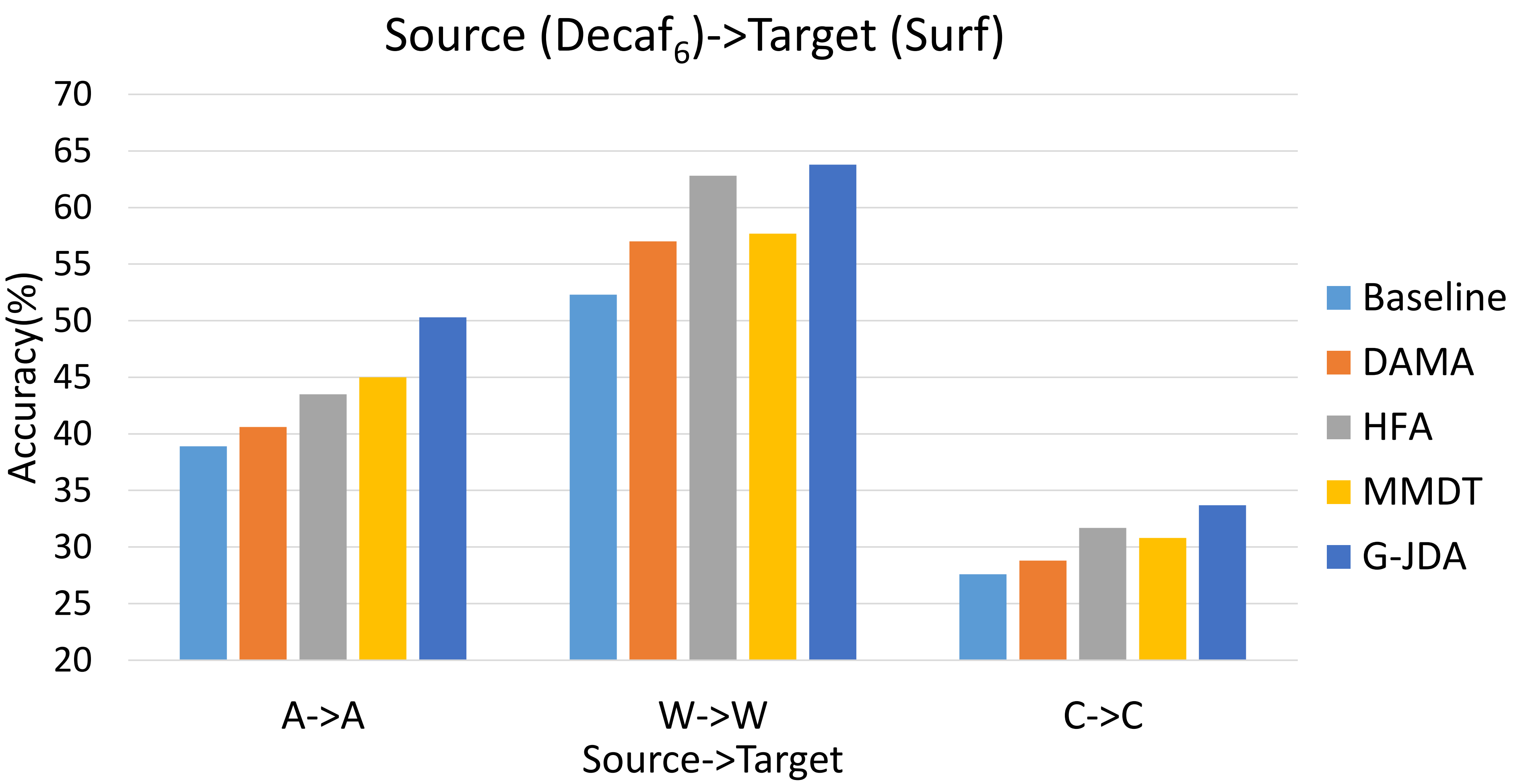
Settings & Features

- 10 overlapped classes across domains
- Feature: Decaf_6 [J. Donahue *et al*., 2014] and Surf [H. Bay *et al*., 2006]
- Source domain : randomly select 20 images per class.
- Target domain: randomly select 3 images per class as labeled data.

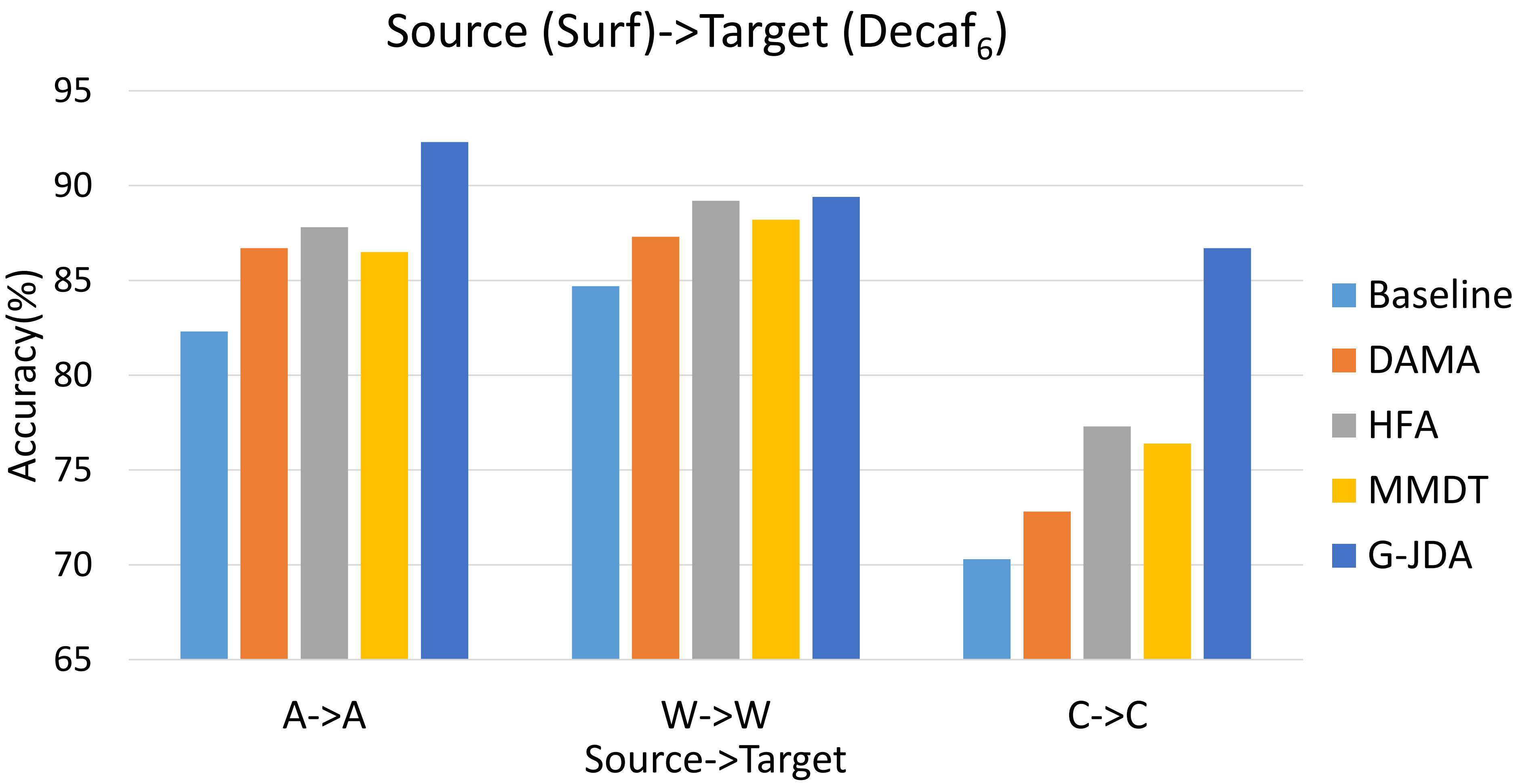
Recent HDA Methods

- Baseline : SVM in target domain [C.-C. Chang and C.-J. Lin, LIBSVM 2011]
- DAMA : [C. Wang and S. Mahadevan, IJCAI 2011]
- HFA : [L. Duan *et al*., ICML 2012]
- MMDT : [J. Hoffman *et al*., ICLR 2013]

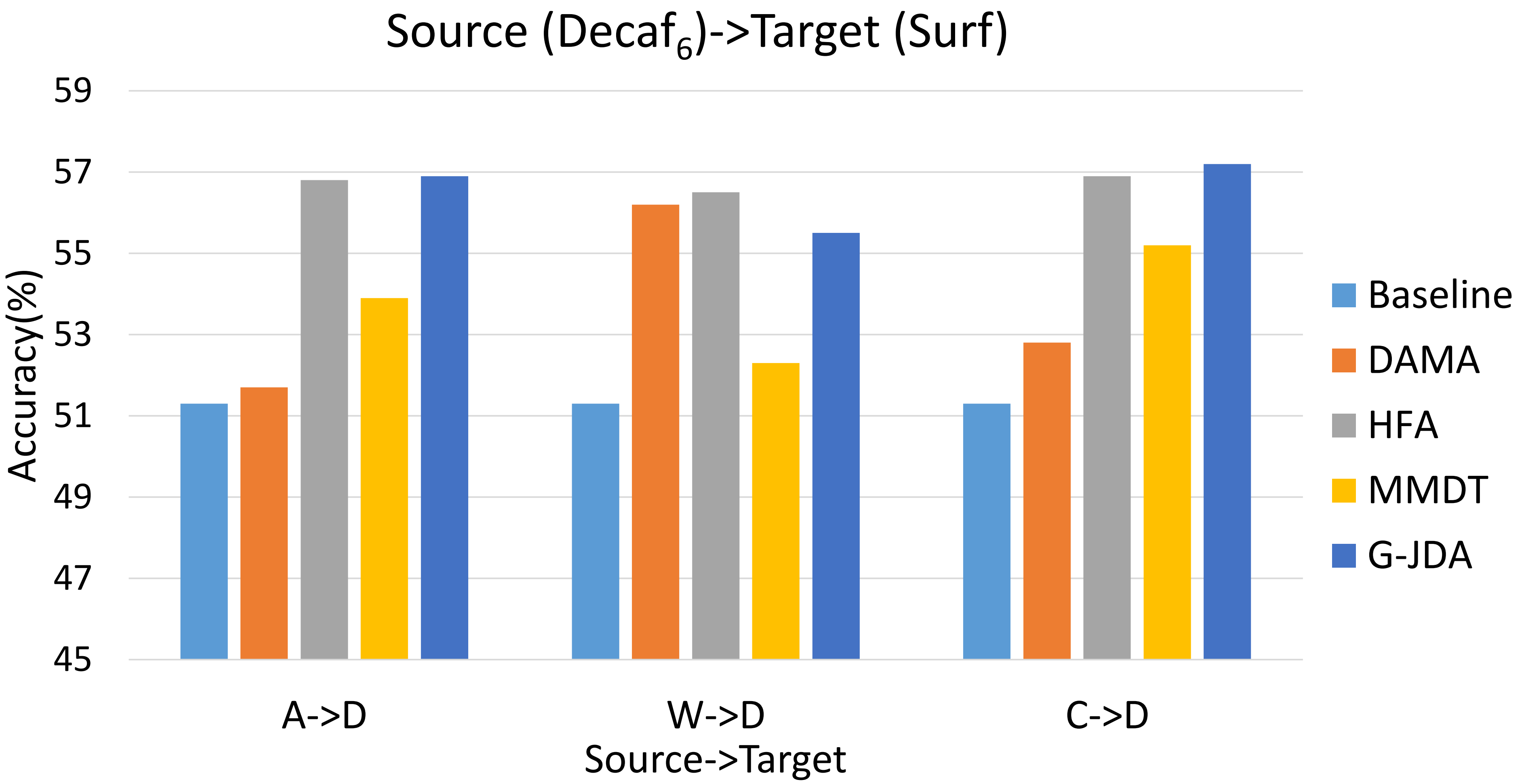
Cross-Feature Recognition



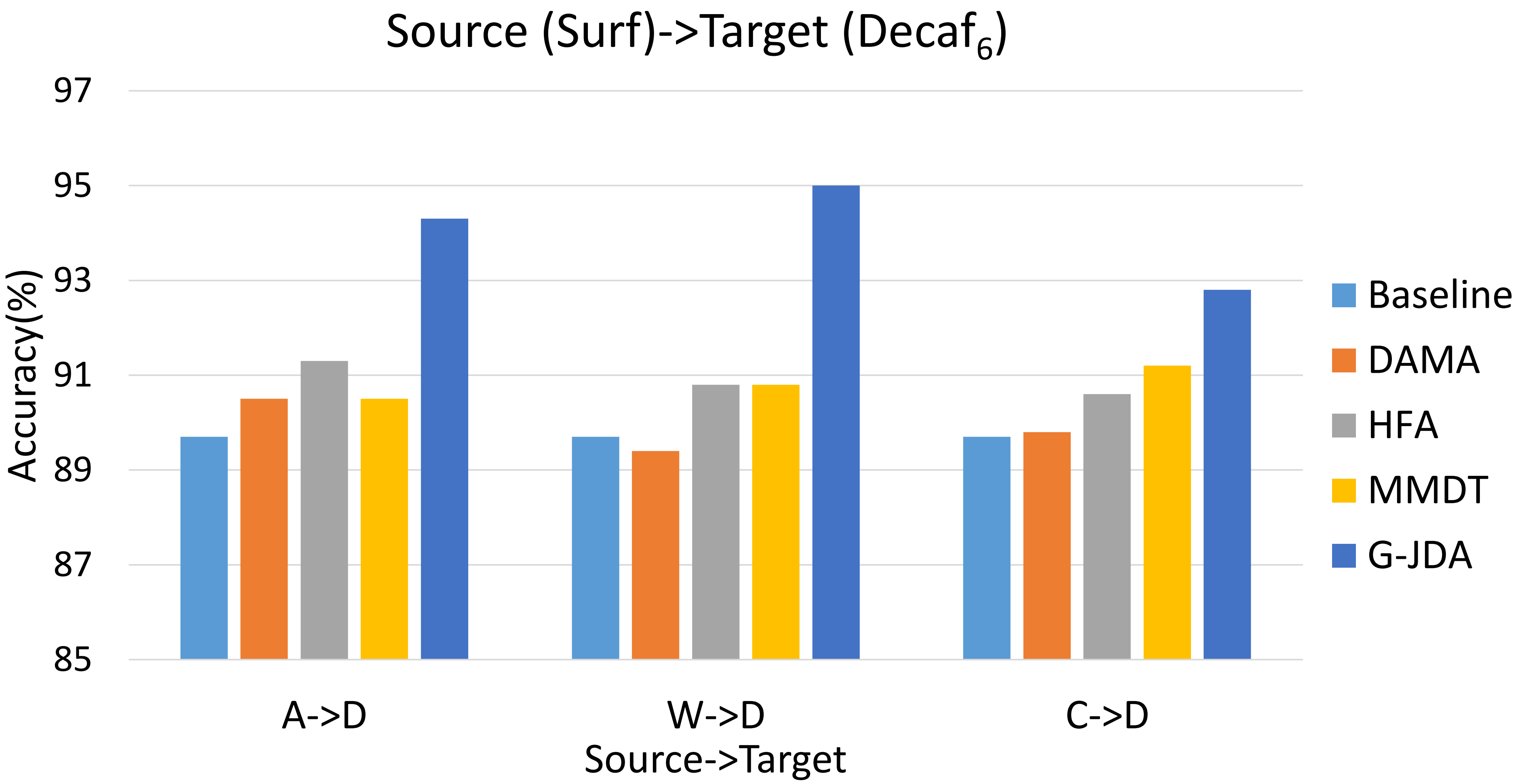
Cross-Feature Recognition (cont'd)



Recognition Across Domains & Features



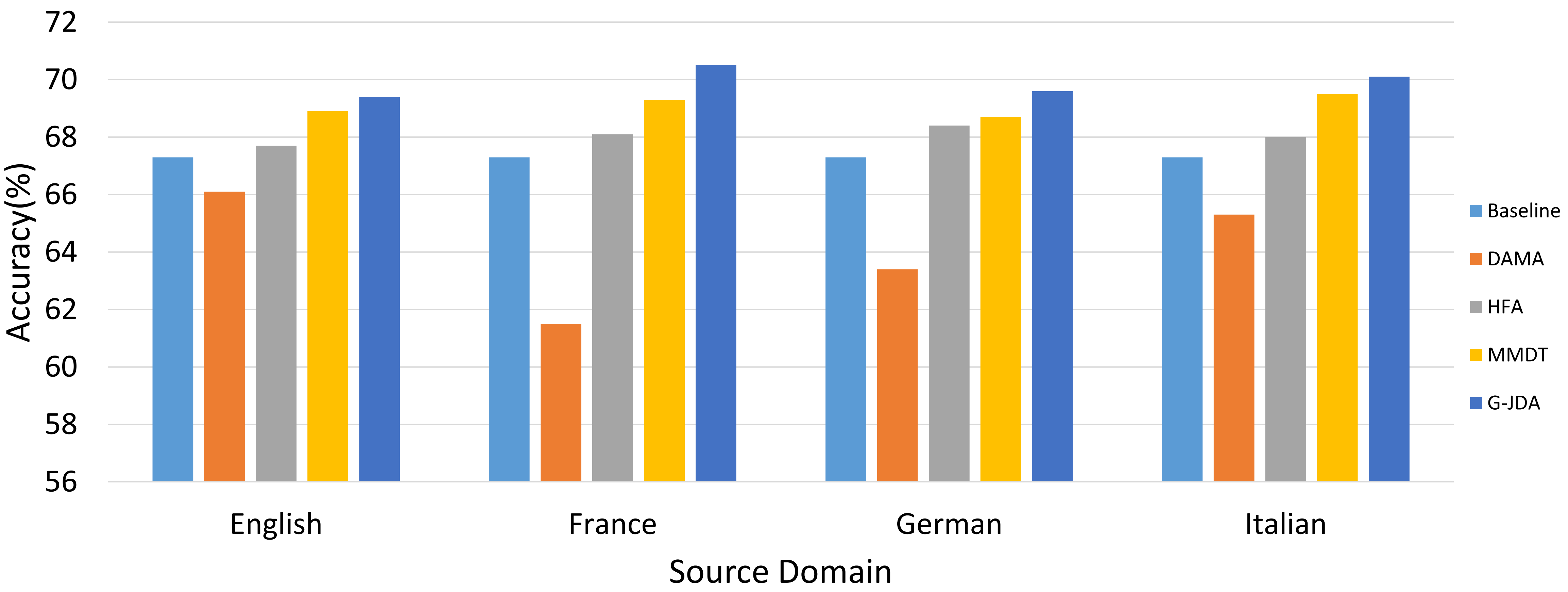
Recognition Across Domains & Features (cont’d)



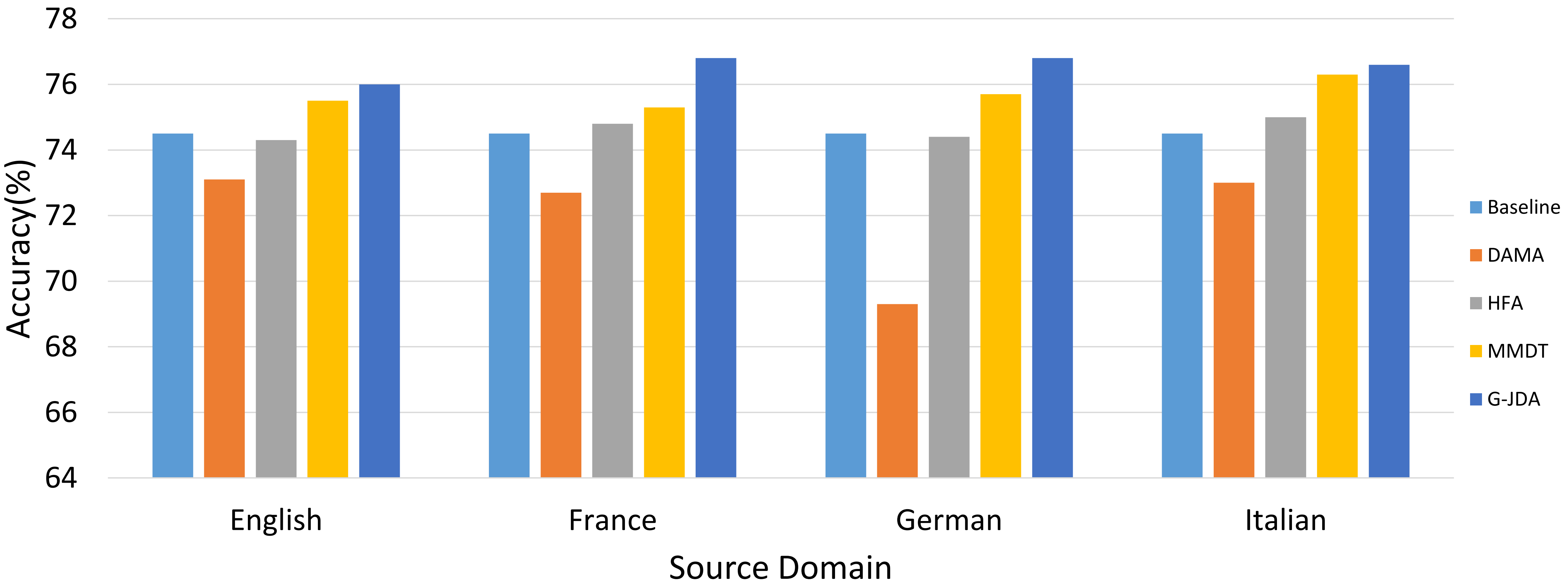
Cross-Lingual Text Categorization

- Multilingual Reuters Collection Dataset [M. Amini *et al.*, NIPS 2009]
- 11K articles from 6 categories in 5 languages
- English, French, Italian, German, and Spanish
- BOW + TF-IDF with 60% energy preserved via PCA [L. Duan *et al.*, ICML ’12]
- Source domain : randomly select 100 articles per class.
- Target domain: randomly select {10,20} articles per class as labeled data, and randomly select 500 articles per class as unlabeled data.

of Labeled Target-Domain (Spanish) Data per Category = 10

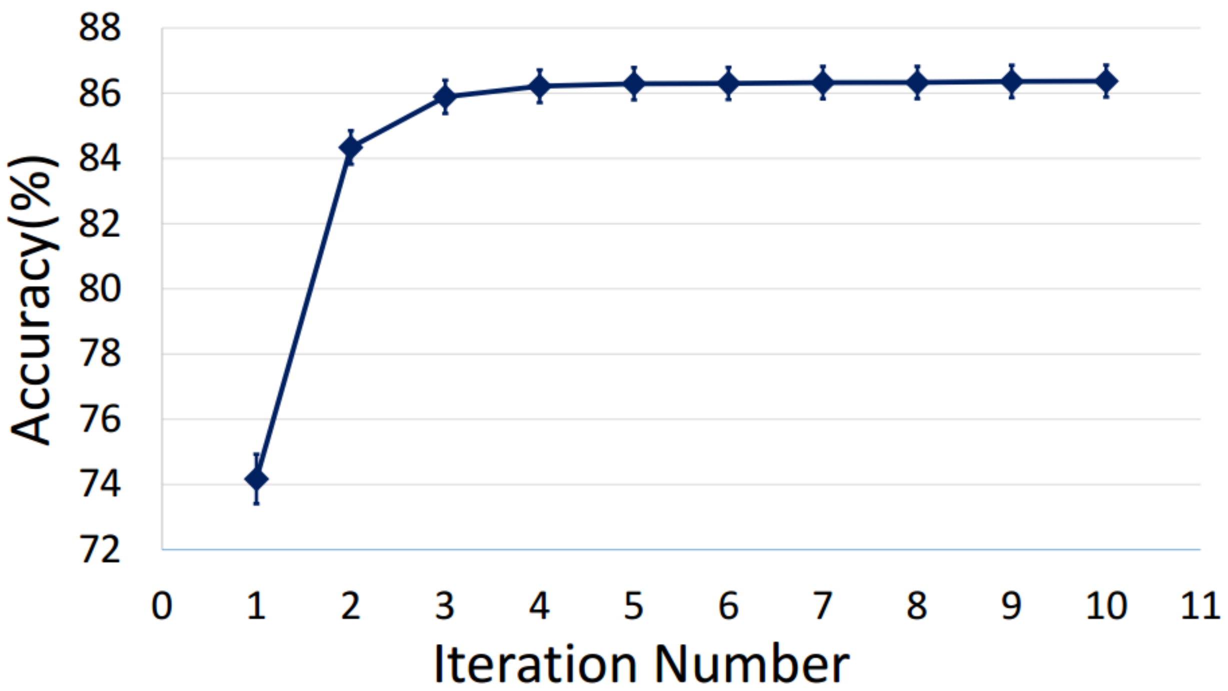


of Labeled Target-Domain (Spanish) Data per Category = 20



Convergence analysis

- Image classification task with cross domains & features
- Source: Caltech with SURF
- Target: DSLR with Decaf_6



Outline

- Motivation
- Domain Adaptation
- Proposed Method
- Experiment
- Conclusion

Conclusion

- We proposed Generalized Joint Distribution Adaptation (G-JDA) for associating & recognizing heterogeneous cross-domain data.

- By learning a pair of feature transformations for source and target-domain data, G-JDA derives a domain-invariant common feature space for addressing the above goal.

- Our experiments on cross-domain visual and text classification tasks verified the effectiveness of our proposed G-JDA for HDA.

Q & A

Thank You!

Outline

- Motivation
- Domain Adaptation
- Proposed Method
- Experiment
- Conclusion
- Appendix

Distribution Measure: MMD Criteria

- Maximum Mean Discrepancy [A. Gretton et al. NIPS 2006]
- MMD is an empirical formula which measures discrepancy between two distributions
- $\{x_1, x_2, \dots, x_n\}$ come from distribution P
- $\{y_1, y_2, \dots, y_m\}$ come from distribution Q
- $MMD(P, Q) = \left\|\frac{1}{n}\sum_i \varphi(x_i) - \frac{1}{m}\sum_j \varphi(y_j)\right\|_H$
- $\varphi(.)$ is kernel function
- We follow *TCA* [S. J. Pan et al. 2011]and *JDA* [M. Long et al. 2013] to choose linear kernel in MMD.

Optimization

- 1. Rewrite the equation as compact trace form

$$E_{mar}(\mathbf{A}_S, \mathbf{A}_T) = \| \frac{1}{n_S} \sum_{i=1}^{n_S} \mathbf{A}_S^\top \mathbf{x}_S^i - \frac{1}{n_T} \sum_{j=1}^{n_T} \mathbf{A}_T^\top \mathbf{x}_T^j \|^2$$

$$E_{cond}^{(c)}(\mathbf{A}_S, \mathbf{A}_T) = \| \frac{1}{n_S^c} \sum_{i=1}^{n_S^c} \mathbf{A}_S^\top \mathbf{x}_S^{i,c} - \frac{1}{n_T^c} \sum_{j=1}^{n_T^c} \mathbf{A}_T^\top \mathbf{x}_T^{j,c} \|^2$$

Optimization

- 1. Rewrite the equation as compact trace form

$$E_{mar} = tr(\mathbf{A}^\top \mathbf{X} \mathbf{M}_0 \mathbf{X}^\top \mathbf{A})$$

$$E_{cond}^{(c)} = tr(\mathbf{A}^\top \mathbf{X} \mathbf{M}_c \mathbf{X}^\top \mathbf{A})$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_S & \mathbf{0}_{d_S \times n_T} \\ \mathbf{0}_{d_T \times n_S} & \mathbf{X}_T \end{pmatrix}$$

$$\mathbf{A} = [\mathbf{A}_S; \mathbf{A}_T]$$

$$(\mathbf{M}_0)_{ij} = \begin{cases} \frac{1}{n_S n_S} & \text{if } i, j \leq n_S \\ \frac{1}{n_T n_T} & \text{if } i, j > n_S \\ \frac{-1}{n_S n_T} & \text{otherwise.} \end{cases}$$

$$(\mathbf{M}_c)_{ij} = \begin{cases} \frac{1}{n_S^c n_S^c} & \text{if } i, j \leq n_S \text{ and } z_i = z_j = c \\ \frac{1}{n_T^c n_T^c} & \text{if } i, j > n_S \text{ and } z_i = z_j = c \\ \frac{-1}{n_S^c n_T^c} & \text{if } \begin{cases} i \leq n_S, j > n_S \\ i > n_S, j \leq n_S \end{cases} \quad \text{and } z_i = z_j = c \\ 0 & \text{otherwise.} \end{cases}$$

Optimization

- 2. The formula becomes

$$\min_{\mathbf{A}=[\mathbf{A}_S;\mathbf{A}_T]} \sum_{i=0}^{C} tr(\mathbf{A}^\top \mathbf{X}\mathbf{M}_i\mathbf{X}^\top \mathbf{A}) + \lambda \left(\|\mathbf{A}_S\|^2 + \|\mathbf{A}_T\|^2\right)$$

$$\text{s.t.} \quad \mathbf{A}^\top \mathbf{X}\mathbf{H}\mathbf{X}^\top \mathbf{A} = \mathbf{I}$$

Optimization

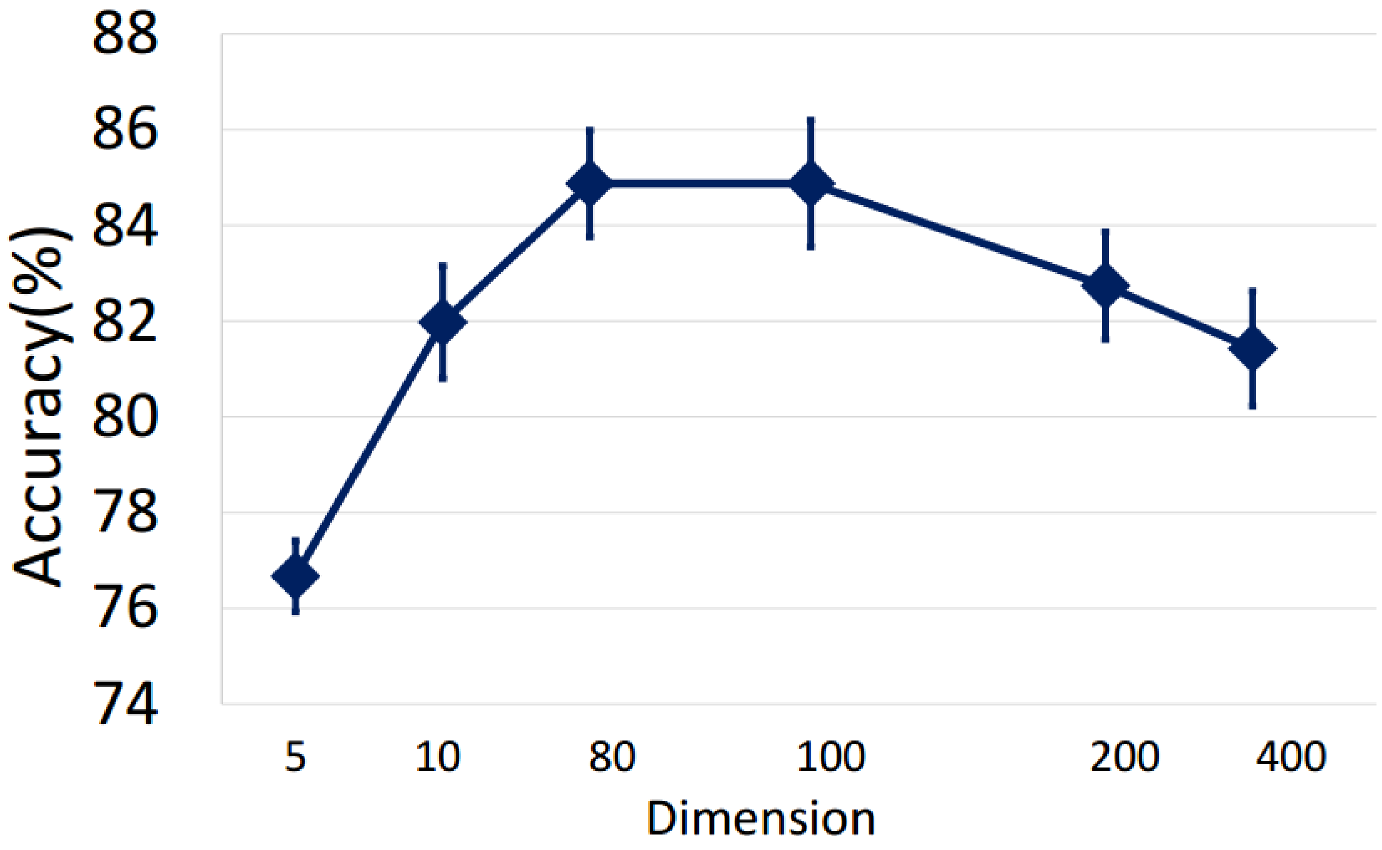
- 3. Solve generalized eigenvalue decomposition problem
- A can be determined by d_K(dimensionality of latent space) smallest eigenvectors

$$(\mathbf{X}\sum_{i=0}^{c}\mathbf{M}_i\mathbf{X}^\top + \mathbf{R})\mathbf{A} = \psi\mathbf{XHX}^\top\mathbf{A}$$

$$\mathbf{R} \in \mathbb{R}^{ns+n_T} = \begin{pmatrix} \lambda\mathbf{I}_{n_S} & \mathbf{0}_{n_S\times n_T} \\ \mathbf{0}_{n_T\times n_S} & \lambda\mathbf{I}_{n_T} \end{pmatrix}$$

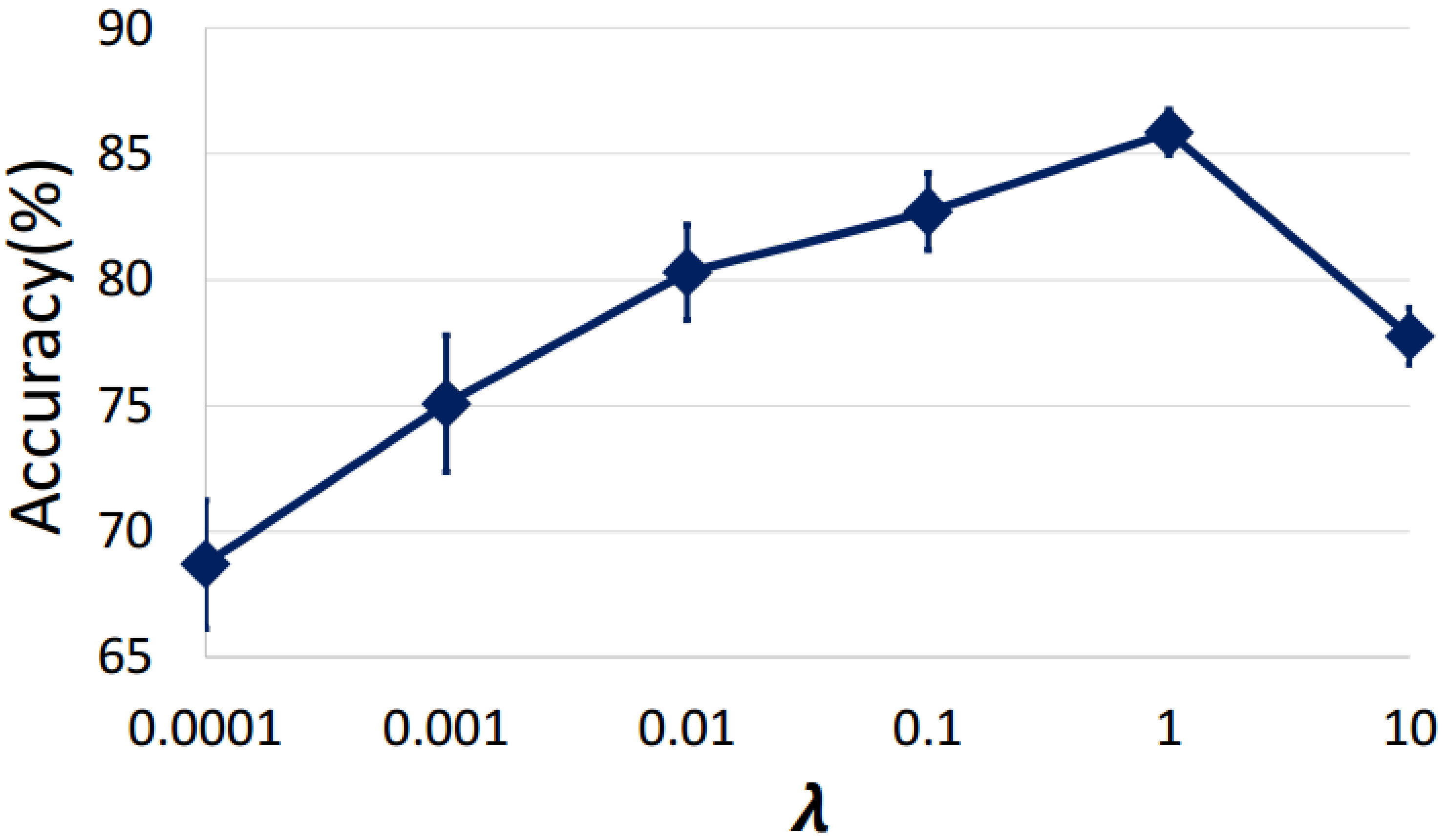
Parameter sensitivity

- Dimensionality of the common feature space



Parameter sensitivity

- Lambda



Parameter sensitivity

- Iteration number