ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

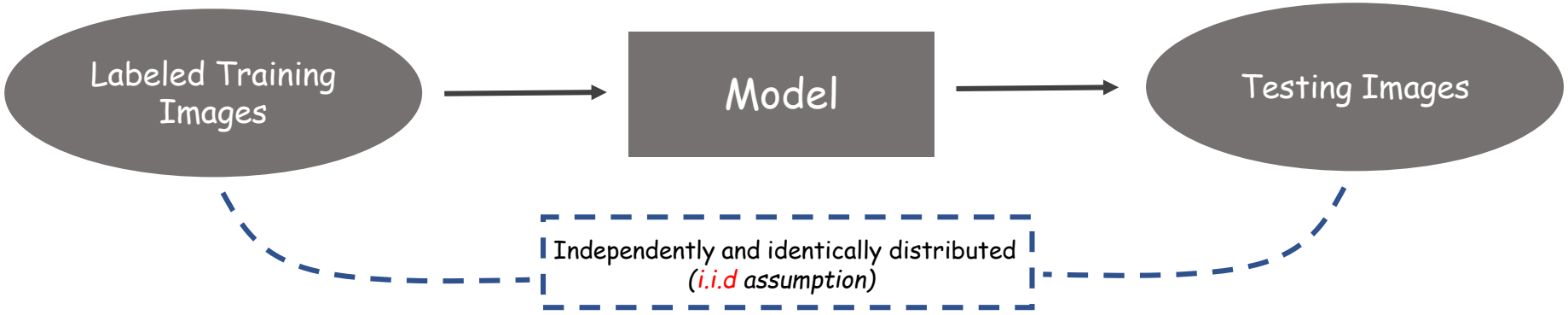
Transferable Attention for Domain Adaptation

Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long,✉ Jianmin Wang
School of Software, Tsinghua University, China
KLiss, MOE; BNRist; Research Center for Big Data, Tsinghua University, China
{wxm17,liliang17,ywr16}@mails.tsinghua.edu.cn {mingsheng,jimwang}@tsinghua.edu.cn

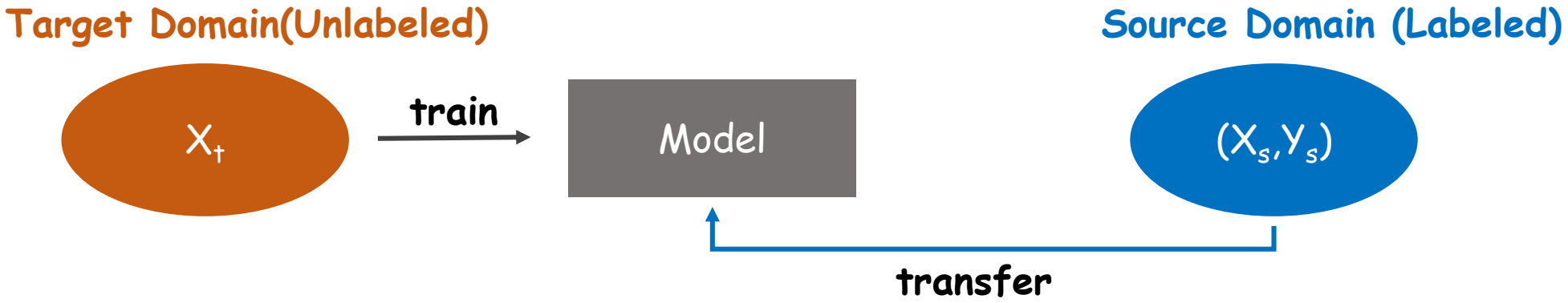
AAAI '19

Unsupervised Domain Adaptation(UDA)

- Supervised Learning



- Unsupervised Domain Adaptation(UDA)



- Domain shift: $P_s(X,Y) \neq P_t(X,Y)$ or $P_s(x_s) \neq P_t(x_t)$
- Feature space and label space are consistent: $x_s, x_t \in \mathbb{R}^d \quad y_s, y_t \in L$

Unsupervised Domain Adaptation(UDA)

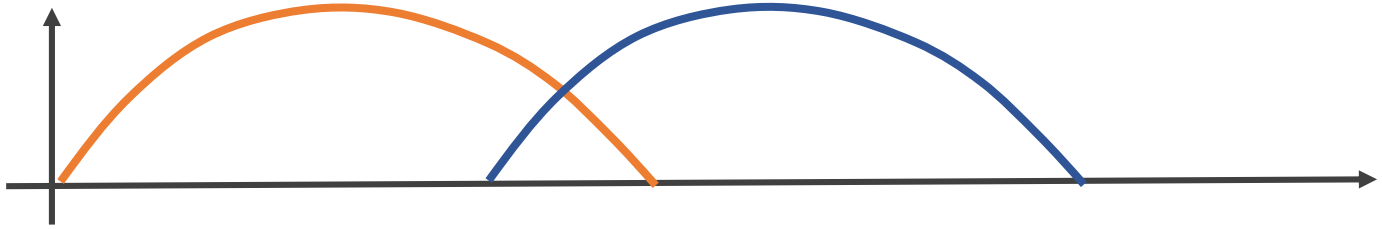
- Unsupervised Domain Adaptation(UDA)

Bike:

Keyboard:

Source domain
with annotations

Target domain
without annotations



Solution: Domain Alignment

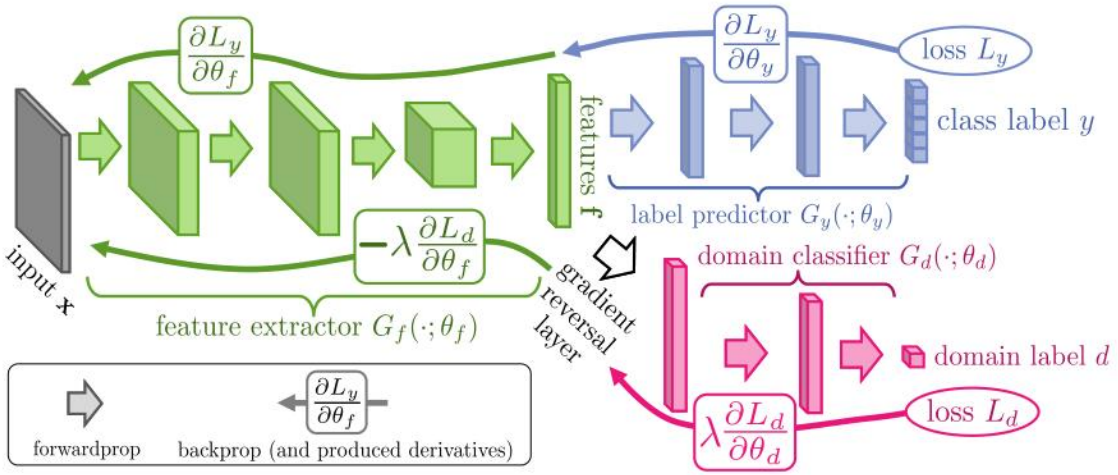
✓ **Goal**: Achieving good performance on the target domain.

Background

✓ 减小最大均值差异 (Maximum Mean Discrepancy)

$$MMD(X,Y) = ||\frac{1}{n}\sum_{i=1}^{n}\phi(x_i) - \frac{1}{m}\sum_{j=1}^{m}\phi(y_j)||_H^2$$

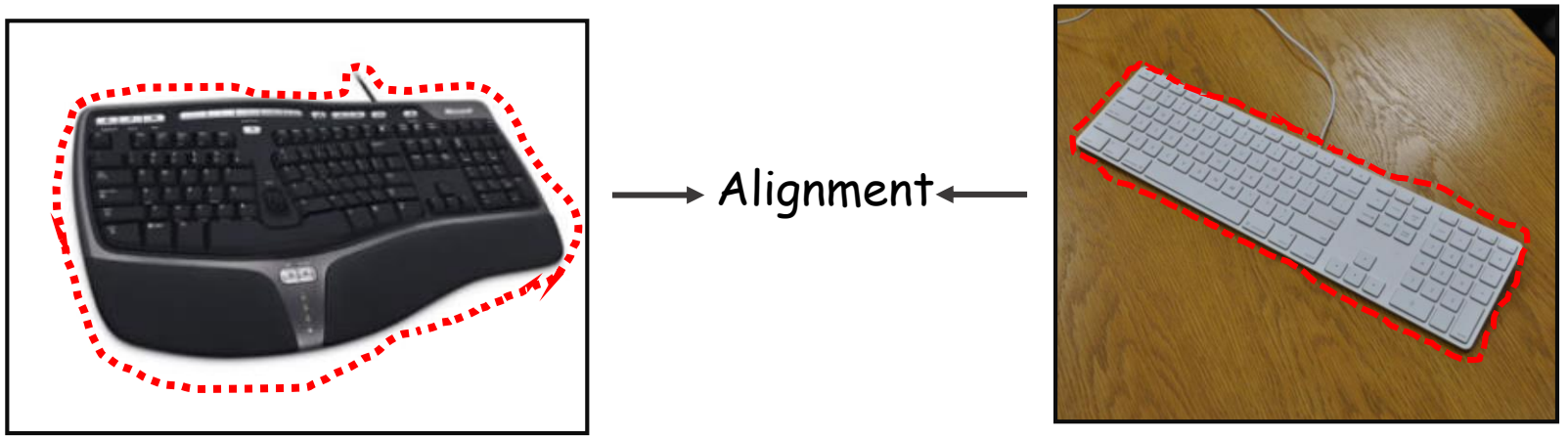
✓ 对抗找到不变特征 (Domain-invariant features)



ICML'15

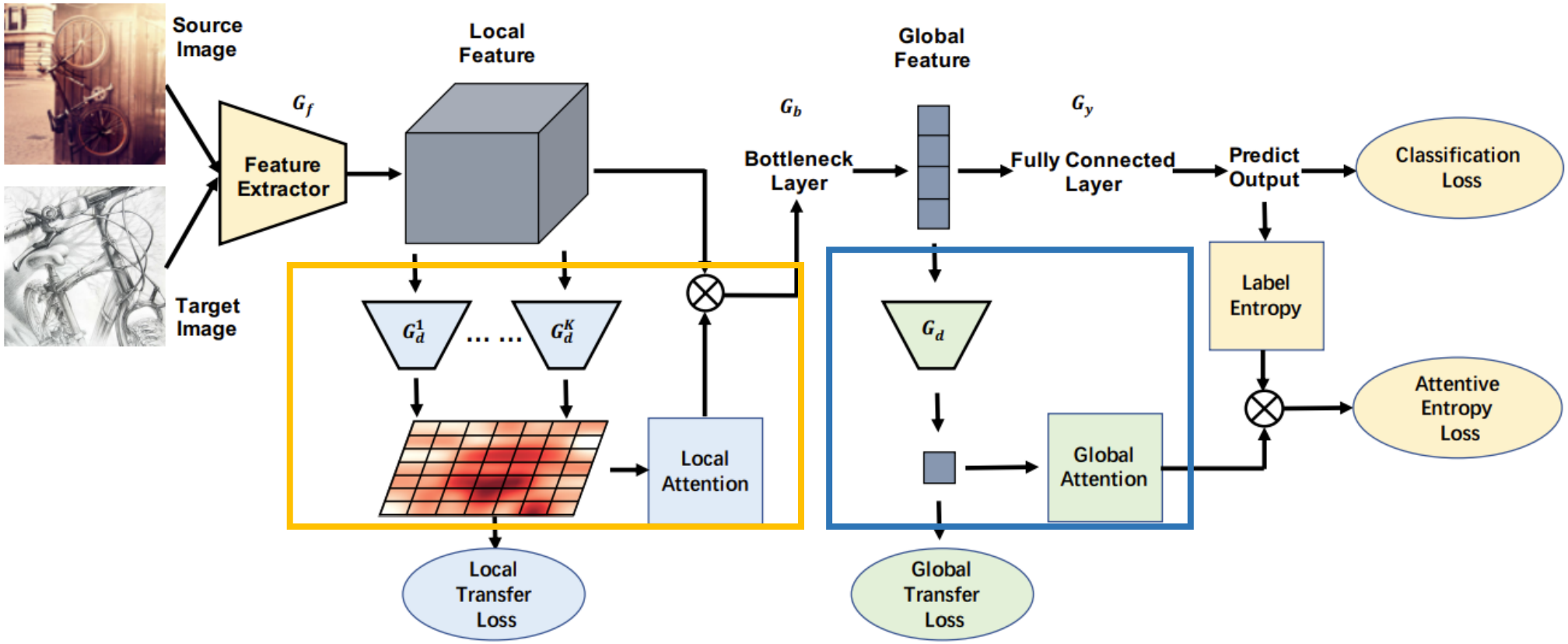
Motivation

- However, it is obvious that not all regions of an image are transferable, while forcefully aligning the untransferable regions may lead to negative transfer.



- Ignore features that are not helpful for category discrimination.

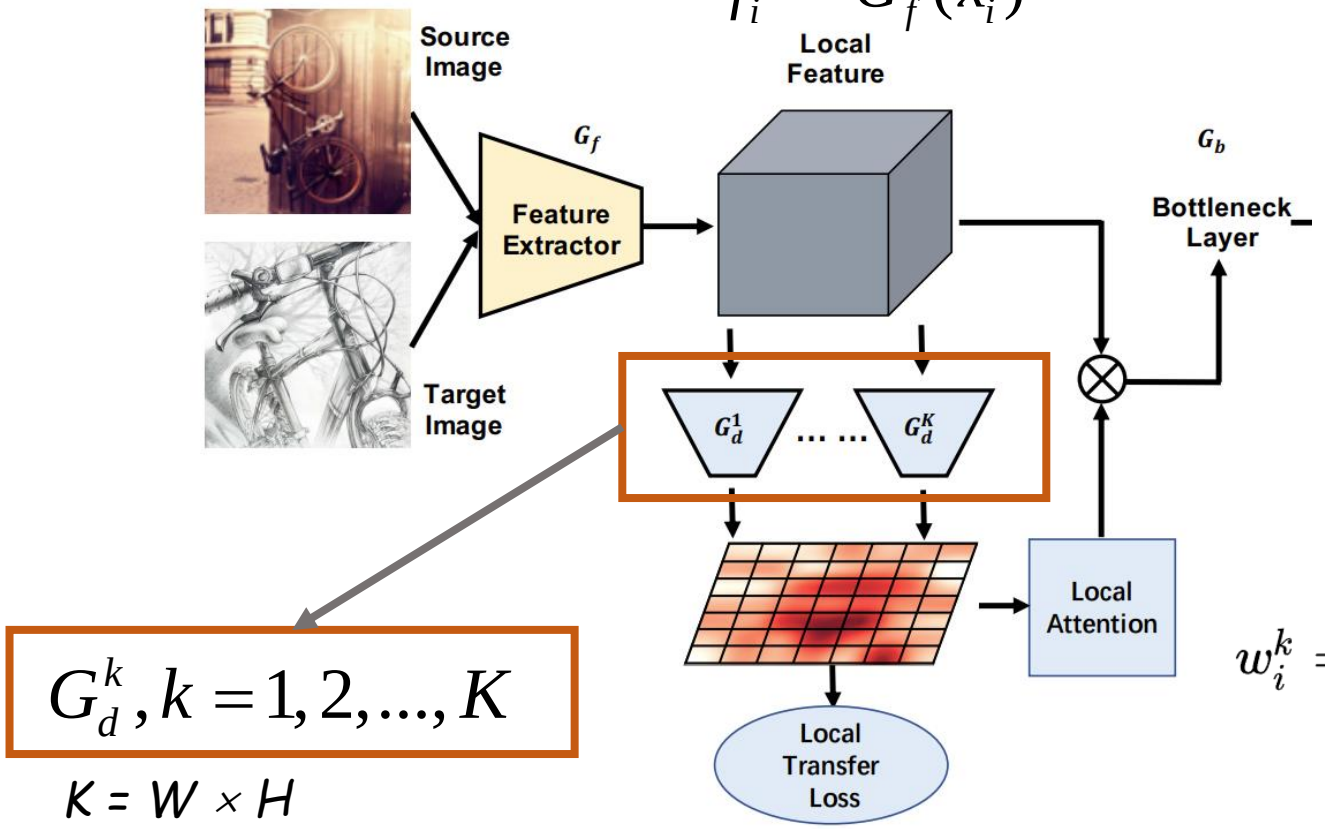
Transferable Attention for Domain Adaptation



- Multi-adversarial network is developed for local attention to highlight the representations of those regions with higher transferability
- Global adversarial network (green) is utilized to enhance the prediction certainty of the images more similar in the feature space across domains.

Transferable Local Attention

$$f_i^k = G_f(x_i)^k \longrightarrow h_i^k = (1 + w_i^k) \cdot f_i^k$$



$$G_d^k, k = 1, 2, ..., K$$

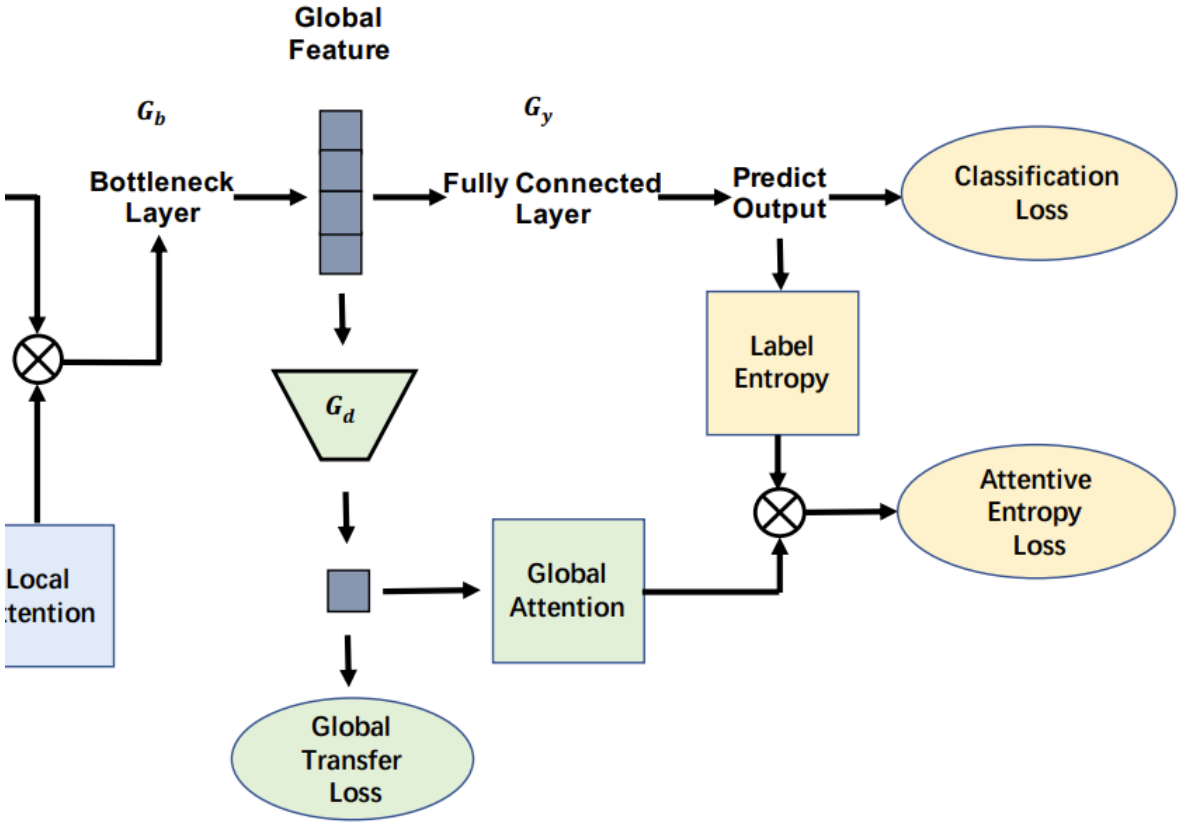
$K = W \times H$

$$H(p) = -\sum_j p_j \cdot \log(p_j)$$

$$w_i^k = 1 - H\left(\hat{d}_i^k\right) = 1 + \sum \hat{d}_i^k \cdot log(\hat{d}_i^k)$$

$$L_l = \frac{1}{Kn} \sum_{k=1}^{K} \sum_{x_i \in \mathcal{D}_s \cup \mathcal{D}_t} L_d(G_d^k(f_i^k), d_i), \quad \hat{d}_i^k = G_d^k(f_i^k)$$

Transferable Global Attention



$$L_y = \frac{1}{n_s} \sum_{\boldsymbol{x}_i \in \mathcal{D}_s} L_y(G_y(G_b(\boldsymbol{h}_i)), y_i), \qquad (8)$$

$$L_h = -\frac{1}{n} \sum_{\boldsymbol{x}_i \in \mathcal{D}_s \cup \mathcal{D}_t} \sum_{j=1}^{c} m_i \cdot \boldsymbol{p}_{i,j} \cdot \log(\boldsymbol{p}_{i,j}), \qquad (7)$$

$$m_i = 1 + H(\hat{d}_i) = 1 - \sum \hat{d}_i \cdot log(\hat{d}_i)$$

Attention Loss

$$L_g = \frac{1}{n} \sum_{\boldsymbol{x}_i \in \mathcal{D}_s \cup \mathcal{D}_t} L_d(G_d(G_b(\boldsymbol{h}_i), d_i)) \quad \hat{d}_i = G_d(G_b(\boldsymbol{h}_i))$$

Finally Objection

$$
\begin{aligned}
C(\theta_f, \theta_b, \theta_y, \theta_d, \theta_d^k|_{k=1}^K) &= L_y + \gamma L_h - \lambda(L_g + L_l) \\
&= \frac{1}{n_s} \sum_{\boldsymbol{x}_i \in \mathcal{D}_s} L_y(G_y(G_b(\boldsymbol{h}_i)), y_i) \\
&- \frac{\gamma}{n} \sum_{\boldsymbol{x}_i \in \mathcal{D}} \sum_{j=1}^{C} m_i \cdot \boldsymbol{p}_{i,j} \cdot \log(\boldsymbol{p}_{i,j}) \\
&- \frac{\lambda}{n} [\sum_{\boldsymbol{x}_i \in \mathcal{D}} L_d(G_d(G_b(\boldsymbol{h}_i), d_i)) \\
&+ \frac{1}{K} \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in \mathcal{D}} L_d(G_d^k((G_f(\boldsymbol{x}_i))^k), d_i)]
\end{aligned}
\tag{9}
$$

minmax

$$
\begin{aligned}
(\hat{\theta}_f, \hat{\theta}_b, \hat{\theta}_y) &= \arg\min_{\theta_f, \theta_b, \theta_y} C\left(\theta_f, \theta_b, \theta_y, \theta_d, \theta_d^k|_{k=1}^K\right), \\
(\hat{\theta}_d, \hat{\theta}_d^1, ..., \hat{\theta}_d^K) &= \arg\max_{\theta_d, \theta_d^1, ..., \theta_d^K} C\left(\theta_f, \theta_b, \theta_y, \theta_d, \theta_d^k|_{k=1}^K\right).
\end{aligned}
\tag{10}
$$

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

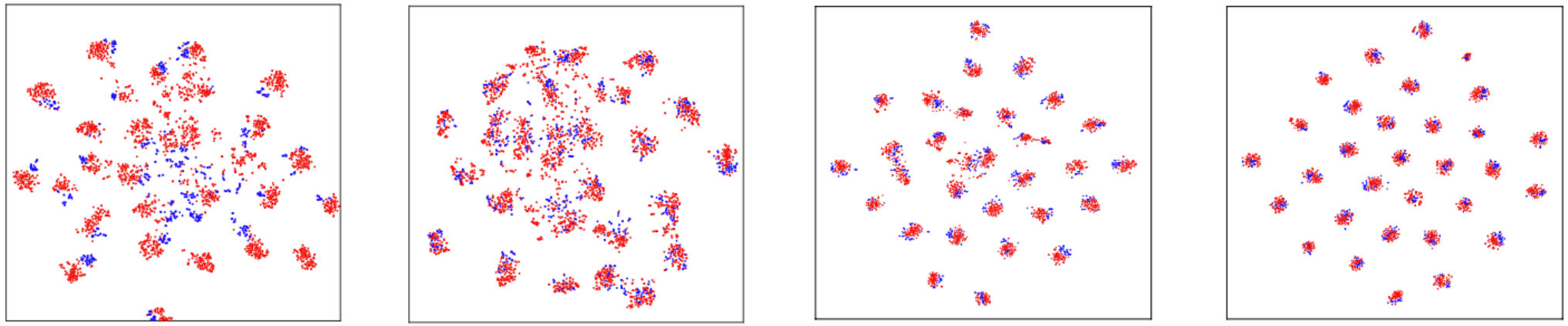
Table 1: Accuracy (%) on *Office-31* for unsupervised domain adaption (ResNet)

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
ResNet-50 (He et al. 2016)	68.4 ± 0.2	96.7 ± 0.1	99.3 ± 0.1	68.9 ± 0.2	62.5 ± 0.3	60.7 ± 0.3	76.1
TCA (Pan et al. 2011)	72.7 ± 0.0	96.7 ± 0.0	99.6 ± 0.0	74.1 ± 0.0	61.7 ± 0.0	60.9 ± 0.0	77.6
GFK (Gong et al. 2012)	72.8 ± 0.0	95.0 ± 0.0	98.2 ± 0.0	74.5 ± 0.0	63.4 ± 0.0	61.0 ± 0.0	77.5
DAN (Long et al. 2015)	80.5 ± 0.4	97.1 ± 0.2	99.6 ± 0.1	78.6 ± 0.2	63.6 ± 0.3	62.8 ± 0.2	80.4
RTN (Long et al. 2016)	84.5 ± 0.2	96.8 ± 0.1	99.4 ± 0.1	77.5 ± 0.3	66.2 ± 0.2	64.8 ± 0.3	81.6
DANN (Ganin et al. 2016)	82.0 ± 0.4	96.9 ± 0.2	99.1 ± 0.1	79.7 ± 0.4	68.2 ± 0.4	67.4 ± 0.5	82.2
ADDA (Tzeng et al. 2017)	86.2 ± 0.5	96.2 ± 0.3	98.4 ± 0.3	77.8 ± 0.3	69.5 ± 0.4	68.9 ± 0.5	82.9
JAN (Long et al. 2017)	85.4 ± 0.3	97.4 ± 0.2	**99.8** ± 0.2	84.7 ± 0.3	68.6 ± 0.3	70.0 ± 0.4	84.3
MADA (Pei et al. 2018)	90.0 ± 0.1	97.4 ± 0.1	99.6 ± 0.1	87.8 ± 0.2	70.3 ± 0.3	66.4 ± 0.3	85.2
SimNet (Pinheiro 2018)	88.6 ± 0.5	98.2 ± 0.2	99.7 ± 0.2	85.3 ± 0.3	**73.4** ± 0.8	71.6 ± 0.6	86.2
GTA (Sankaranarayanan et al. 2018)	89.5 ± 0.5	97.9 ± 0.3	99.8 ± 0.4	87.7 ± 0.5	72.8 ± 0.3	71.4 ± 0.4	86.5
TADA (local)	89.4 ± 0.4	**98.7** ± 0.2	99.8 ± 0.2	87.2 ± 0.2	66.4 ± 0.2	65.3 ± 0.3	84.5
TADA (global)	92.9 ± 0.4	98.2 ± 0.2	99.8 ± 0.2	88.9 ± 0.2	69.6 ± 0.2	71.0 ± 0.3	86.7
TADA (local+global)	**94.3** ± 0.3	**98.7** ± 0.1	**99.8** ± 0.2	**91.6** ± 0.3	72.9 ± 0.2	**73.0** ± 0.3	**88.4**

Table 2: Accuracy (%) on *Office-Home* for unsupervised domain adaption (ResNet)

Method	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
ResNet-50 (He et al. 2016)	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
DAN (Long et al. 2015)	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN (Ganin et al. 2016)	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN (Long et al. 2017)	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
TADA (local)	47.3	69.1	75.2	56.9	66.4	69.1	55.9	46.9	75.7	68.2	56.2	80.4	63.9
TADA (global)	51.3	66.0	76.5	58.6	69.3	70.3	58.3	52.0	77.1	70.2	57.0	81.5	65.7
TADA (local+global)	**53.1**	**72.3**	**77.2**	**59.1**	**71.2**	**72.1**	**59.7**	**53.1**	**78.4**	**72.4**	**60.0**	**82.9**	**67.6**

Experiments



(a) ResNet (b) DANN (c) MADA (d) TADA

Figure 2: The t-SNE visualization of features learned by (a) ResNet, (b) DANN, (c) MADA, and (d) TADA (red: **A**; blue: **W**).

Experiments

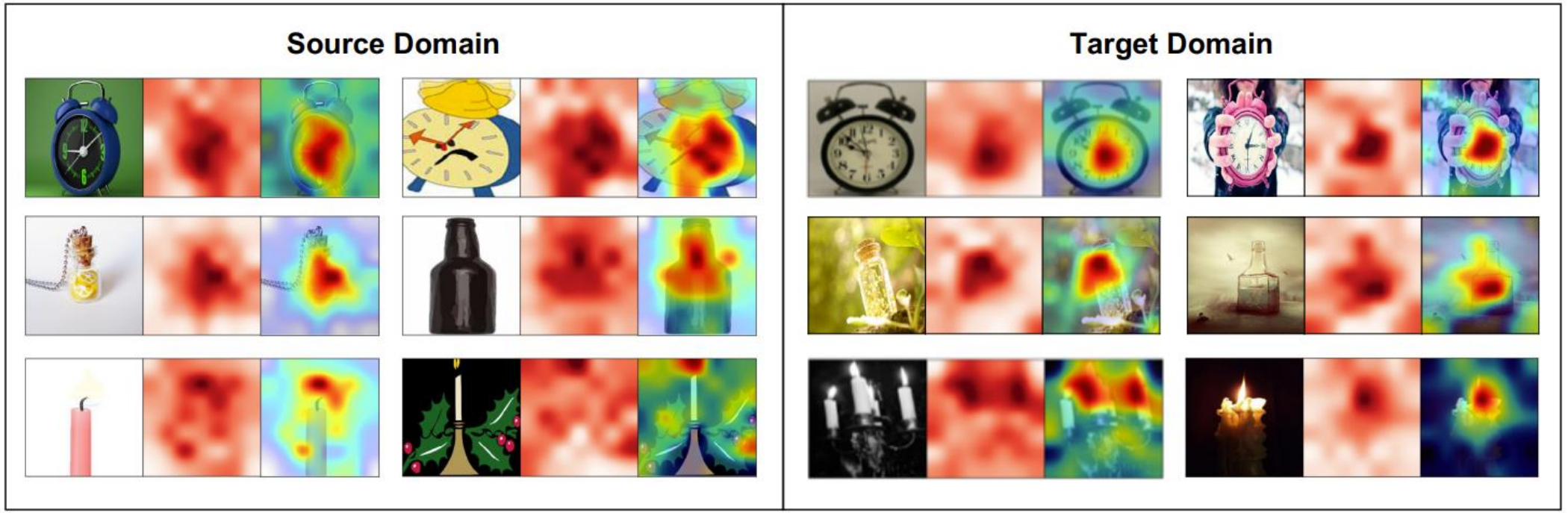


Figure 3: Attention visualization of the last convolutional layer of ResNet on *Office-Home*. The images on the left are randomly sampled from source domain (**Ar**) while the right from target domain (**Rw**). In each group of images, the original input images, the corresponding attentions and the attentions shown in the original input images are illustrated from left to right respectively.

Thanks