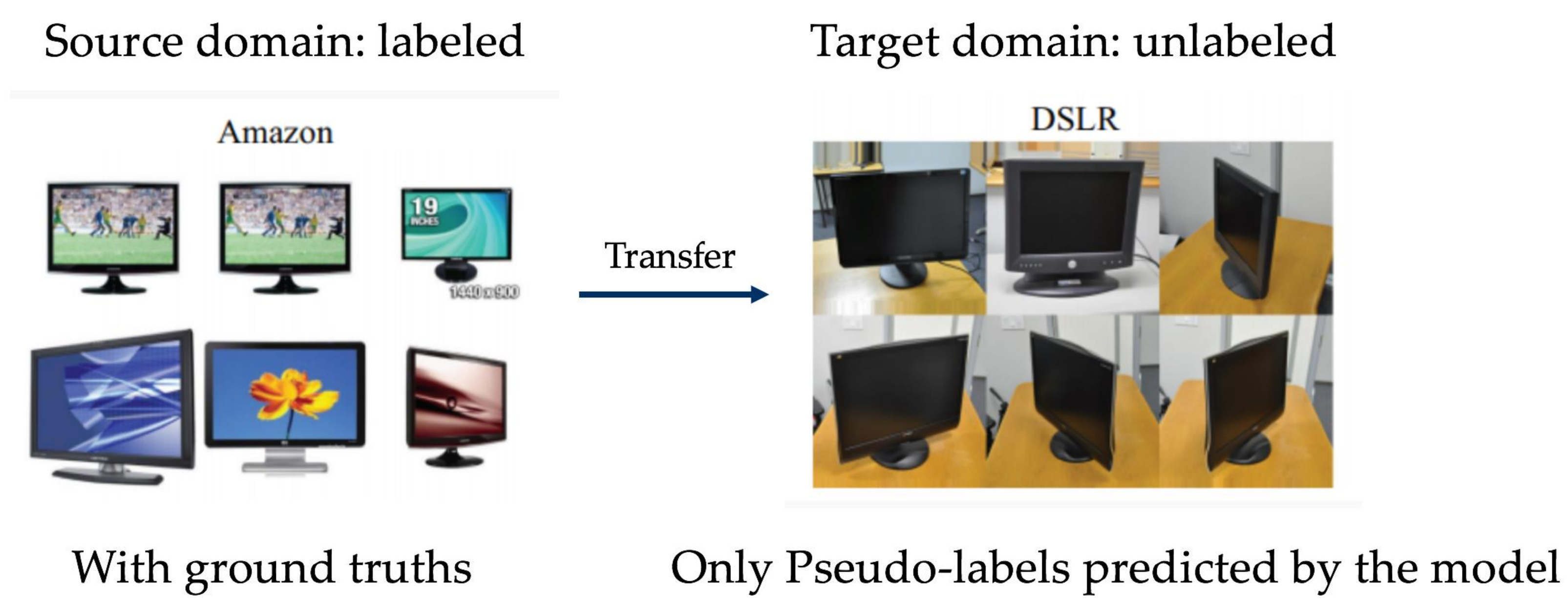
Adversarial-Learned Loss for Domain Adaptation

Minghao Chen, Shuai Zhao, Deng Cai, Haifeng Liu

State Key Lab of CAD&CG, College of Computer Science, Zhejiang University, Hangzhou, China;
Fabu Inc., Hangzhou, China;
Alibaba-Zhejiang University Joint Institute of Frontier Technologies, Hangzhou, China

Domain Adaptation Task

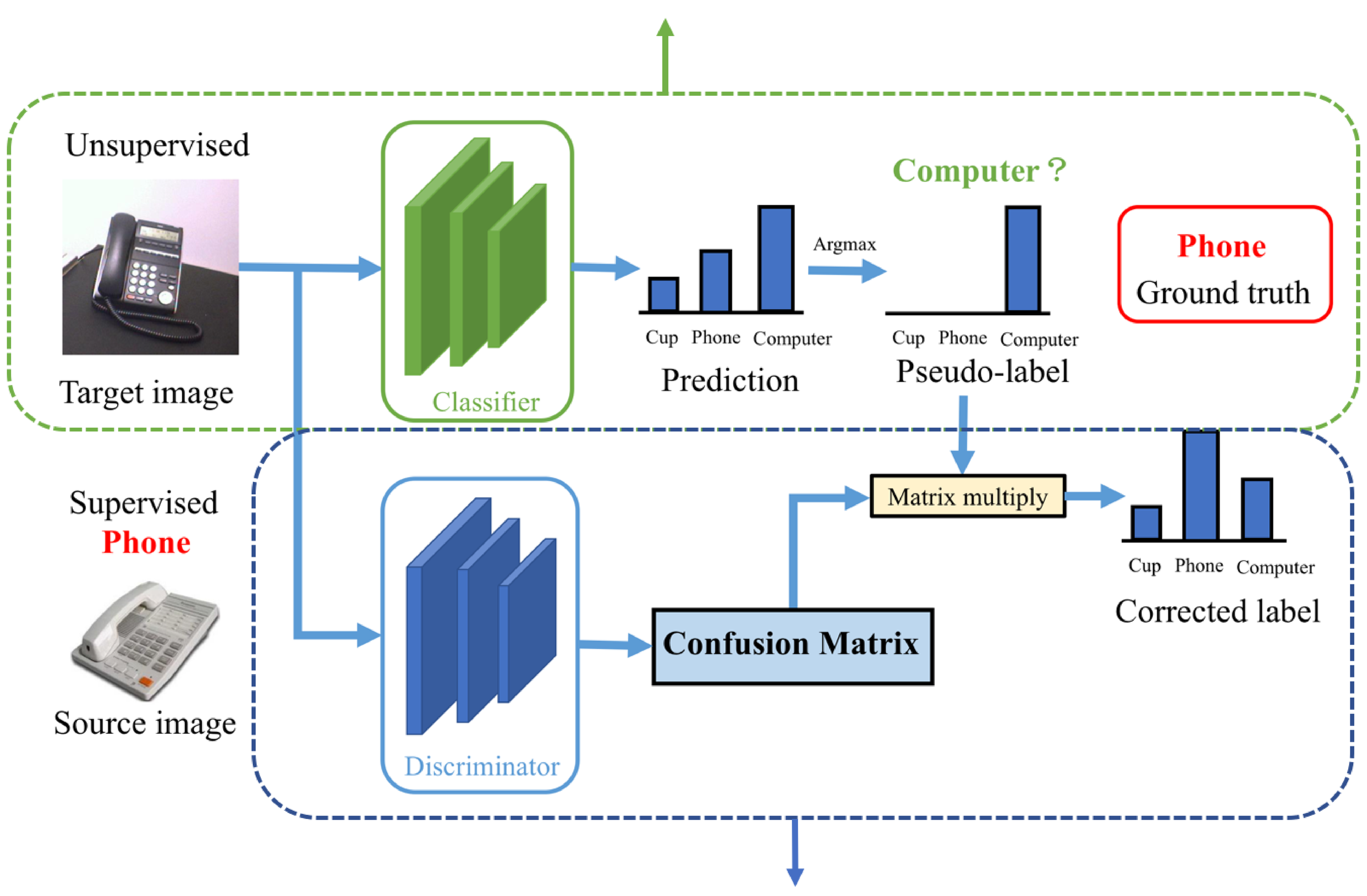
Unsupervised Domain Adaptation (UDA) :
We have labeled data on the source domain and unlabeled data on the target domain. We want to classify the target samples utilizing the source data.



Pseudo-labels: we train the model on the source and use pseudo-labels predicted by the model as the training label on the target.

Motivation

- Pseudo-labels might be incorrect and contain noise.



- We can use a discriminator producing a **confusion matrix** to correct the noise in pseudo-labels.
- The discriminator is train by **Noise-correcting Domain Discrimination,** a kind of class-aware domain adversarial learning.

Confusion matrix:

measure the difference between ground truth and pseudo-label.

$$\mathcal{L}_T(x_t, \mathcal{L}) = \sum_{k=1}^{K} p(y_t = k|x_t)\mathcal{L}(\mathbf{p}_t, k)$$

$$= \sum_{k=1}^{K}\sum_{l=1}^{K} p(y_t = k|\hat{y}_t = l, x_t)p(\hat{y}_t = l|x_t)\mathcal{L}(\mathbf{p}_t, k)$$

$$= \sum_{k=1}^{K}\sum_{l=1}^{K} \eta_{kl}^{(x_t)} p(\hat{y}_t = l|x_t)\mathcal{L}(\mathbf{p}_t, k),$$

where $\eta^{(x_t)}$ is the confusion matrix, $\hat{y}_t$ is the pseudo-labels, and $\mathcal{L}(\mathrm{p}_t, \mathrm{k})$ is a basic loss, e.g. cross entropy loss, unhinge loss.

Methods

To simplify the noisy label problem, we assume that **the noise is class-wise uniform with vector $\boldsymbol{\eta}^{(x_t)}$:**

Definition 1. Noise is *class-wise uniform* with vector $\xi^{(x_t)} \in \mathbb{R}^K$, if $\eta_{kl}^{(x_t)} = \xi_k^{(x_t)}$ for $k = l$, and $\eta_{kl}^{(x_t)} = \frac{1-\xi_l^{(x_t)}}{K-1}$ for $k \neq l$.

We propose to use a discriminator to learn the vector $\xi^{(x_t)}$.

Noise-correcting Domain Discrimination

- Correct pseudo-labels to ground truth for source data:

$$\mathcal{L}_{Adv}(x_s, y_s) = \mathcal{L}_{BCE}(\mathbf{c}^{(x_s)}, \mathbf{y}_s)$$

- Correct pseudo-labels to the opposite distribution for target data:

$$\mathcal{L}_{Adv}(x_t) = \mathcal{L}_{BCE}(\mathbf{c}^{(x_t)}, \mathbf{u}^{(\hat{y}_t)}).$$

Domain adversarial learning the generator and the discriminator.

$$\max_{G} \min_{D} E_{(x_s,y_s),x_t} \mathcal{L}_{Adv}(x_s, y_s, x_t)$$

Corrected Pseudo-labels:

$$\mathcal{L}_T(x_t, \mathcal{L}_{unh}) = \sum_{k,l} \eta_{kl}^{(x_t)} p(\hat{y}_t = l|x_t)\mathcal{L}_{unh}(\mathbf{p}_t, k)$$

$$= \sum_{k} \mathbf{c}_k^{(x_t)} \mathcal{L}_{unh}(\mathbf{p}_t, k).$$

Quantitative Results

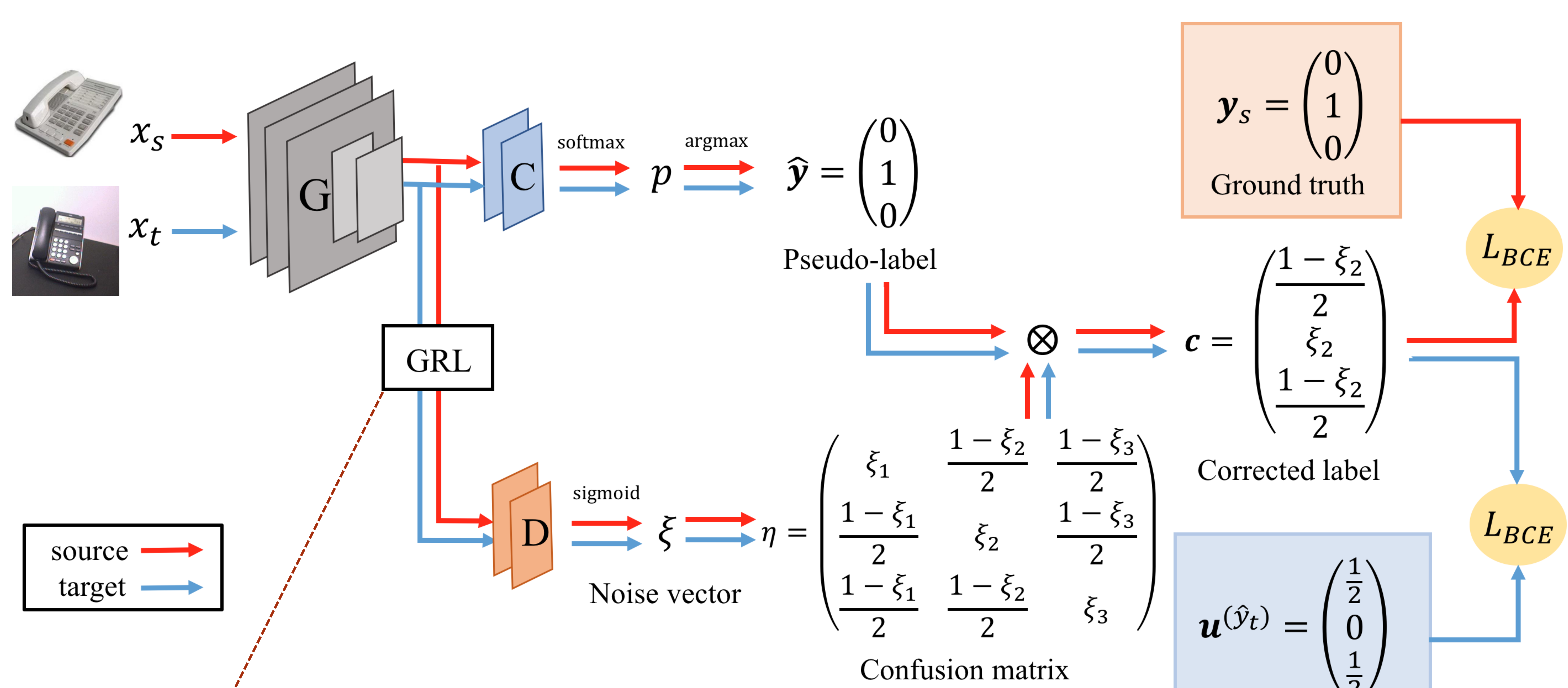
Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
ResNet-50 (He et al. 2016)	68.4±0.2	96.7±0.1	99.3±0.1	68.9±0.2	62.5±0.3	60.7±0.3	76.1
DANN (Ganin et al. 2016)	82.0±0.4	96.9±0.2	99.1±0.1	79.7±0.4	68.2±0.4	67.4±0.5	82.2
ADDA (Tzeng et al. 2017)	86.2±0.5	96.2±0.3	98.4±0.3	77.8±0.3	69.5±0.4	68.9±0.5	82.9
JAN (Long et al. 2017b)	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
CBST (Zou et al. 2018)	87.8±0.8	98.5±0.1	**100±0.0**	86.5±1.0	71.2±0.4	70.9±0.7	85.8
CAN (Zhang et al. 2018)	92.5	**98.8**	**100.0**	90.1	72.1	69.9	87.2
CDAN+E (Long et al. 2017a)	94.1±0.1	98.6±0.1	**100.0±0.0**	92.9±0.2	71.0±0.3	69.3±0.3	87.7
MCS (Liang et al. 2019)	-	-	-	-	-	-	87.8
ALDA	**95.6±0.5**	97.7±0.1	**100.0±0.0**	**94.0±0.4**	**72.2±0.4**	**72.5±0.2**	**88.7**

Table 1: Accuracy (%) of different unsupervised domain adaptation methods on Office-31 (ResNet-50)

Method	U → M	M → U	S → M	Avg
Sourceonly	77.5±0.8	82.0±1.2	66.5±1.9	75.3
DANN (Ganin et al. 2016)	74.0	91.1	73.9	79.7
ADDA (Tzeng et al. 2017)	90.1	89.4	76.0	85.2
CDAN+E (Long et al. 2017a)	98.0	95.6	89.2	94.3
MT+CT (French, Mackiewicz, and Fisher 2018)	92.3±8.6	88.1±0.34	93.3±5.8	91.2
MCD (Saito et al. 2018)	94.1±0.3	96.5±0.3	96.2±0.4	95.6
MCS (Liang et al. 2019)	98.2	**97.8**	91.7	95.9
ALDA ($\delta = 0.9$)	98.1±0.2	94.8±0.1	95.6±0.6	96.2
ALDA ($\delta = 0.8$)	98.2±0.1	95.4±0.4	97.5±0.3	97.0
ALDA ($\delta = 0.6$)	**98.6±0.1**	95.6±0.3	**98.7±0.2**	**97.6**
ALDA ($\delta = 0.0$)	98.4±0.2	95.0±0.1	97.0±0.2	96.8
Targetonly	99.5±0.0	97.3±0.2	99.6±0.1	98.8

δ is the threshold for pseudo-labels.

Digits datasets: USPS to MNIST (U → M), MNIST to USPS (M → U), and SVHN to MNIST (S → M).



GRL is a gradient reverse layer to achieve the minimax optimization.