# Enlarging Discriminative Power by Adding an Extra Class in Unsupervised Domain Adaptation

Hai H. Tran, Sumyeong Ahn, Taeyoung Lee, and Yung Yi

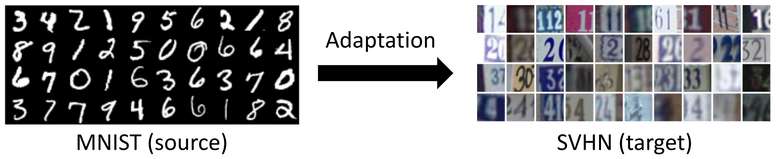
KAIST

LANADA

## Introduction

### Domain Adaptation

- Source data distribution ≠ target data distribution, but **same task**
- Goal: learn a well-performing model **on target data**



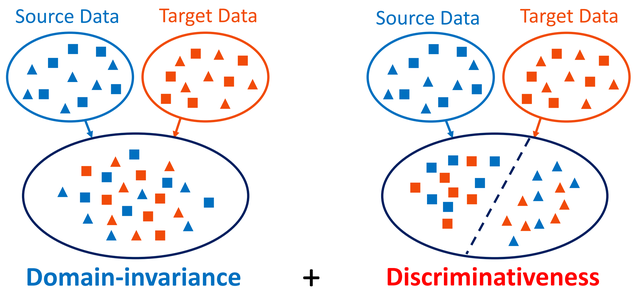

### Unsupervised Domain Adaptation

- Source data: all samples are labeled
- Target data: **no labels** at all

## Background

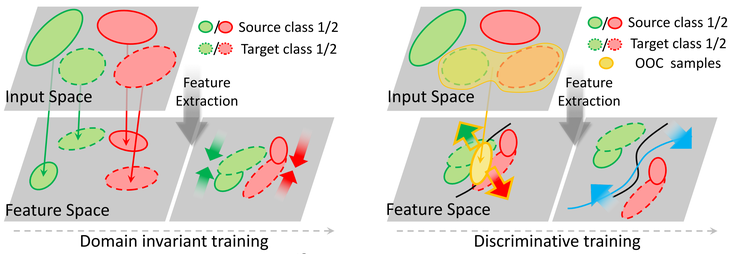
### Key Requirements for a Successful Adaptation

- Domain-invariance: be consistent in different domains
- Discriminativeness: be easy to classify



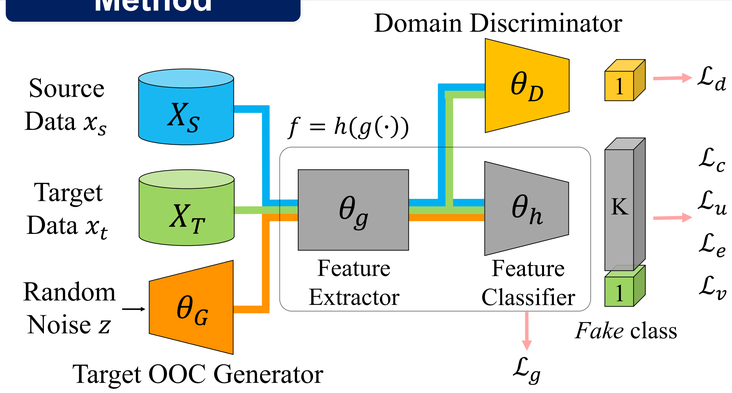

## Key Idea

Improving discriminativeness by generating out-of-class samples, and classifying them to an extra $(K+1)^{\text{th}}$ class.




OOC samples and $(K+1)^{\text{th}}$ class **increase the distance** between real clusters, which helps the classifier **place the decision boundary easier**.

## Method




$g(\cdot)$ **Feature Extractor** extracts common features of source and target domain

$h(\cdot)$ **Feature Classifier** classifies the extracted features from $g(\cdot)$

$D(\cdot)$ **Domain Discriminator** indicates if an input is from source or target domain

$G(\cdot)$ **Generator** generates out-of-class (OOC) samples

**Domain-invariance training:** fooling $D$ to make $g$ generate domain-invariant features

*Source Cross Entropy*
$$\mathcal{L}_c(\theta;\mathcal{D}_S) = \mathbb{E}_{x,y\sim\mathcal{D}_S}\left[\log P_\theta(\hat{y}=y|x, y\le K)\right]$$

*Discriminator Cross Entropy*
$$\mathcal{L}_d(\theta_g,\theta_D;\mathcal{D}_S,\mathcal{D}_T) = \mathbb{E}_{x\sim\mathcal{X}_S}\left[\log D(g(x))\right] + \mathbb{E}_{x\sim\mathcal{X}_T}\left[\log(1-D(g(x)))\right],$$

*OPTIMIZATION*
$$\max_{\theta}\min_{\theta_D}\left[\mathcal{L}_c(\theta;\mathcal{D}_S) + \lambda_d\mathcal{L}_d(\theta_g,\theta_D;\mathcal{X}_S,\mathcal{X}_T)\right]$$

**Improving discriminativeness:** classifying OOC samples to improve the classification power

*Unsupervised Loss*
$$\mathcal{L}_u(\theta;\mathcal{X}_T,P_z) = \mathbb{E}_{x\sim\mathcal{X}_T}\left[\log P_\theta(\hat{y}\le K|x)\right] + \mathbb{E}_{z\sim P_z}\left[\log P_\theta(\hat{y}=K+1|G(z))\right]$$

*Generator FM Loss*
$$\mathcal{L}_g(\theta_G;\mathcal{X}_T,P_z) = \left\|\mathbb{E}_{x\sim\mathcal{X}_T}\left[\phi(x)\right] - \mathbb{E}_{z\sim P_z}\left[\phi(G(z))\right]\right\|$$

*Entropy Loss*
$$\mathcal{L}_e(\theta;\mathcal{D}_T) = -\mathbb{E}_{x\sim\mathcal{D}_T}\left[f(x)^\top \ln f(x)\right]$$

*Virtual Adversarial Training*
$$\mathcal{L}_v(\theta;\mathcal{D}) = \mathbb{E}_{x\sim\mathcal{D}}\left[\max_{\|r\|\le\epsilon} \mathrm{D}_{\mathrm{KL}}(f(x)\,\|\,f(x+r))\right]$$

*OPTIMIZATION*
$$\max_{\theta}\ \mathcal{L}_c(\theta;\mathcal{D}_S) + \lambda_u\mathcal{L}_u(\theta;\mathcal{X}_T,P_z) + \lambda_s\mathcal{L}_v(\theta;\mathcal{D}_S) + \lambda_t\left[\mathcal{L}_v(\theta;\mathcal{D}_T) + \mathcal{L}_e(\theta;\mathcal{D}_T)\right],$$
$$\min_{\theta_G}\ \mathcal{L}_g(\theta_G;\mathcal{X}_T,P_z),$$

**GADA Training Algorithm:** alternative training with 3 steps

$$\max_{\theta}\min_{\theta_D}\min_{\theta_G} \underbrace{\mathcal{L}_c(\theta;\mathcal{D}_S)}_{\text{(a)}} + \underbrace{\lambda_d\mathcal{L}_d(\theta_g,\theta_D;\mathcal{X}_S,\mathcal{X}_T)}_{\text{(b)}} + \underbrace{\lambda_s\mathcal{L}_v(\theta;\mathcal{D}_S) + \lambda_t\left[\mathcal{L}_v(\theta;\mathcal{D}_T) + \mathcal{L}_e(\theta;\mathcal{D}_T)\right]}_{\text{(c)}} + \underbrace{\lambda_u\mathcal{L}_u(\theta;\mathcal{X}_T,P_z) + \mathcal{L}_g(\theta_G;\mathcal{X}_T,P_z)}_{\text{(c)}}$$

(a) achieves classification power from source data

(b) improves domain-invariance

(c) improves discriminative extraction power

## Evaluation

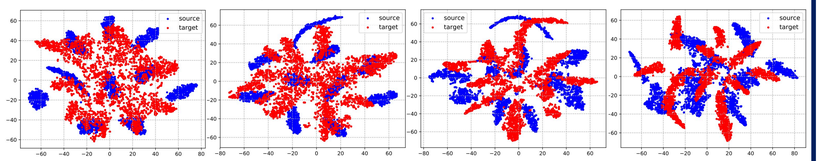
### Domain Adaptation Tasks

| Source / Target | MNIST / SVHN | SVHN / MNIST | MNIST / MNIST-M | DIGITS / SVHN | CIFAR / STL | STL / CIFAR |
|---|---|---|---|---|---|---|
| DANN[8] | 35.7 | 71.1 | 81.5 | 90.3 | - | - |
| DSN[10] | - | 82.7 | 83.2 | 91.2 | - | - |
| ATT[19] | 52.8 | 86.2 | 94.2 | 92.9 | - | - |
| MCD[26] | - | 96.2 | - | - | - | - |
| DA Assoc.[27] | - | 97.6 | 89.5 | 91.9 | - | - |
| GAGL[20] | 74.6 | 96.7 | 94.9 | 93.1 | 77.0 | 61.5 |
| Without instance-normalized input | | | | | | |
| VADA[14] | 47.5 | 97.9 | 97.7 | 94.8 | 80.0 | 73.5 |
| CoDA[13] | 55.3 | 98.8 | 99.0 | **96.1** | **81.4** | **76.4** |
| GADA* (ours) | 69.9 | 98.2 | 98.2 | 94.9 | 79.7 | 73.0 |
| GADA (ours) | **72.3** | **98.9** | **99.1** | 95.7 | 80.5 | 75.1 |
| VADA+DIRT-T[14] | 54.5 | 99.4 | 98.9 | 96.1 | - | 75.3 |
| CoDA+DIRT-T[13] | 63.0 | 99.4 | 99.1 | 96.5 | - | **77.6** |
| GADA*+DIRT-T (ours) | 82.4 | 99.4 | 98.9 | 96.3 | - | 75.3 |
| GADA+DIRT-T (ours) | **83.7** | **99.6** | **99.3** | **96.6** | - | 76.5 |
| With instance-normalized input | | | | | | |
| VADA[14] | 73.3 | 94.5 | 95.7 | 94.9 | 78.3 | 71.4 |
| CoDA[13] | 81.7 | 98.7 | 98.0 | 96.0 | **80.6** | 74.7 |
| GADA* (ours) | 78.7 | **99.0** | 97.2 | 95.3 | 78.9 | 72.1 |
| GADA (ours) | **83.6** | 98.8 | **98.2** | **96.1** | 80.1 | **74.9** |
| VADA+DIRT-T[14] | 76.5 | 99.4 | 98.7 | 96.2 | - | 73.3 |
| CoDA+DIRT-T[13] | 88.0 | 99.4 | 98.8 | 96.5 | - | 75.9 |
| GADA*+DIRT-T (ours) | 84.5 | 99.4 | 98.8 | 96.5 | - | 74.7 |
| GADA+DIRT-T (ours) | **90.0** | **99.6** | **99.0** | **96.8** | - | **76.2** |

### Generated samples

Original MNIST | Generated MNIST | Original SVHN | Generated SVHN

### Feature space comparison in MNIST → SVHN



VADA (acc: 70.6%) | GADA (acc: 83.6%) | VADA+DIRT-T (acc: 75.75%) | GADA+DIRT-T (acc: 90%)

### Ablation study

| | $\mathcal{L}_c$ | $\mathcal{L}_d$ | $\mathcal{L}_e$ | $\mathcal{L}_v$ | $\mathcal{L}_u, \mathcal{L}_g$ | MNIST → SVHN |
|---|---|---|---|---|---|---|
| DANN(ours) | ✓ | ✓ | | | | 66.3 |
| DANN+$\mathcal{L}_e$ | ✓ | ✓ | ✓ | | | 68.1 (+1.8) |
| DANN+$\mathcal{L}_v$ | ✓ | ✓ | | ✓ | | 69.9 (+3.6) |
| GADA* | ✓ | ✓ | | | ✓ | 78.7 (+12.4) |
| VADA(ours) | ✓ | ✓ | ✓ | ✓ | | 70.6 (+4.3) |
| GADA | ✓ | ✓ | ✓ | ✓ | ✓ | **83.6 (+17.3)** |