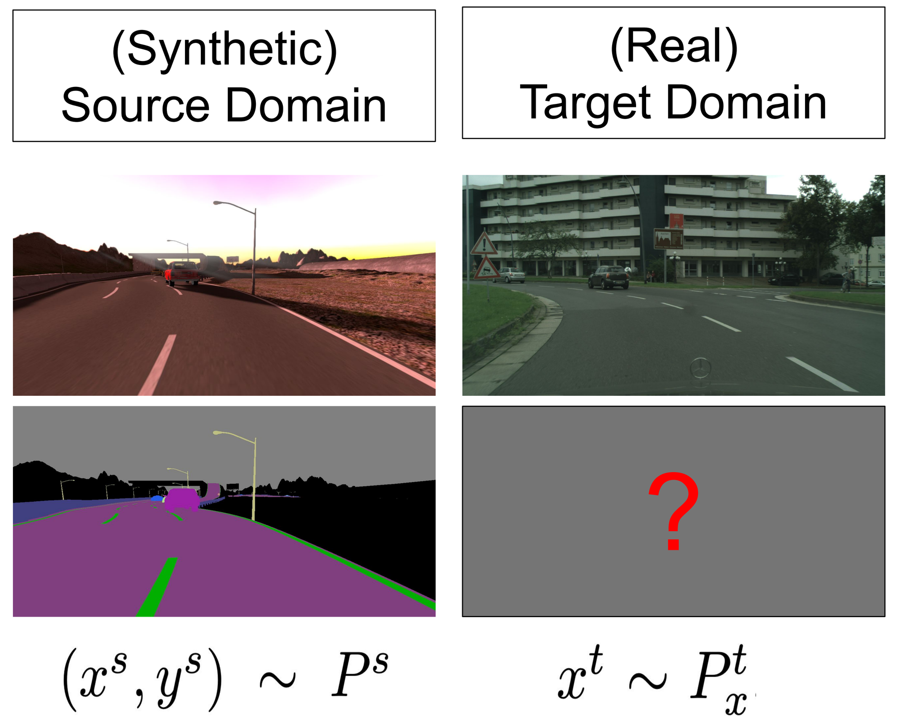
# Unsupervised Domain Adaptation via Regularized Conditional Alignment

Safa Cicek and Stefano Soatto

## Motivation



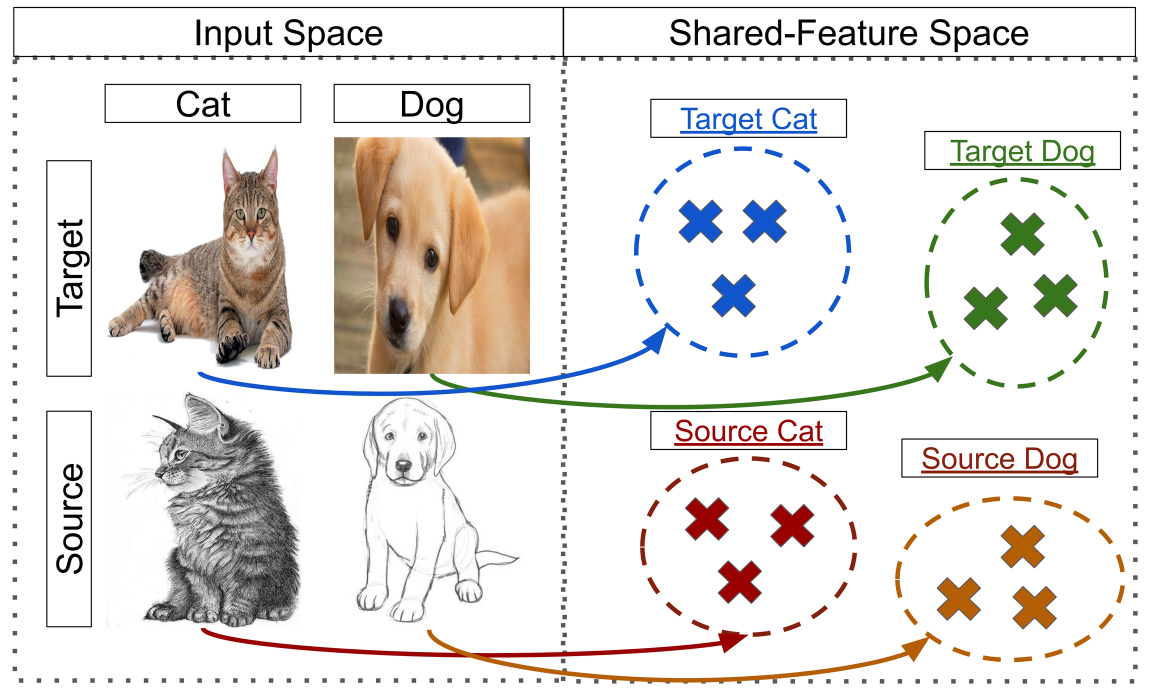

$(x^s, y^s) \sim P^s$ $\quad x^t \sim P^t_x$

$KL(P^s || P^t) > 0$

$\min_f E_{(x,y)\sim P^t} \ell_{CE}(f(x); y)$

$f: X \to Y$

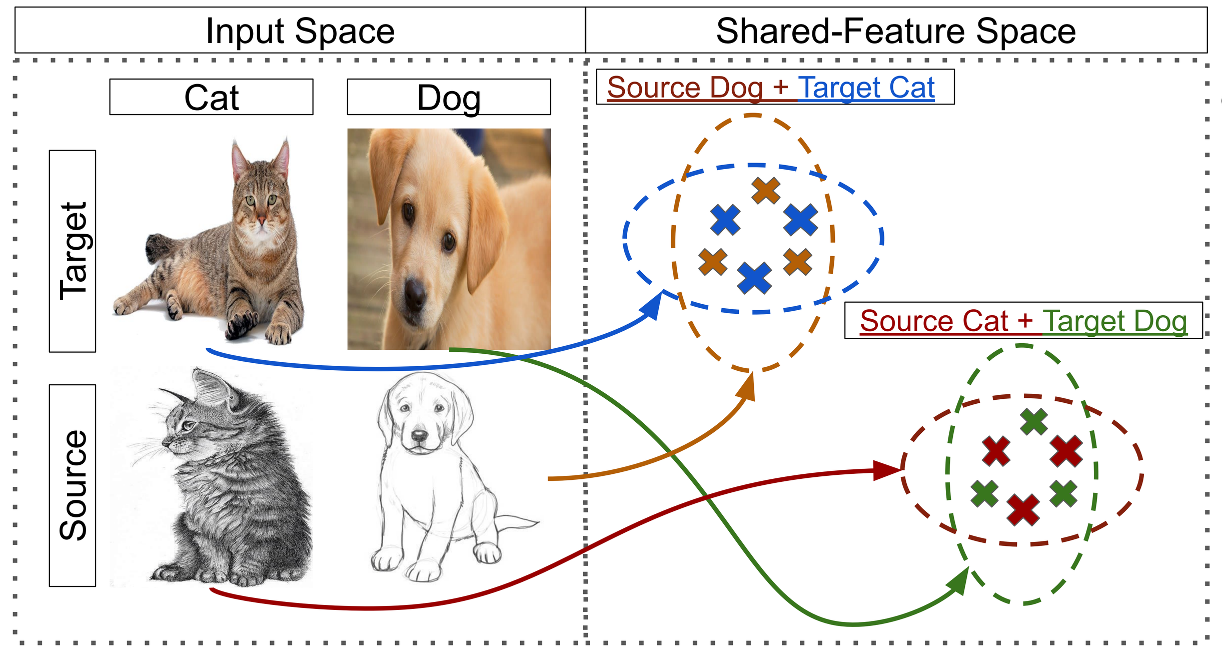
## Shared-Feature Space for UDA




- Moment matching between source and target features (e.g. MMD) [1,2]

[1] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474, 2014.
[2] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. arXiv preprint arXiv:1502.02791, 2015.

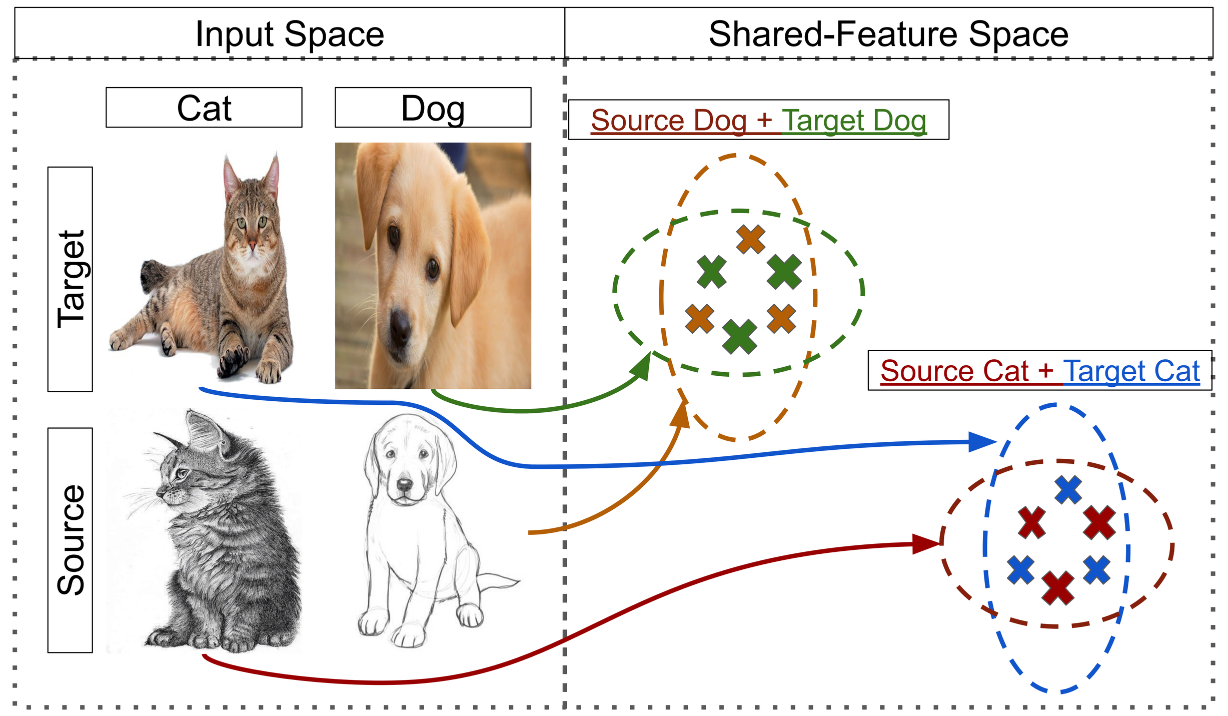
## DANN Aligns Marginal Distributions!



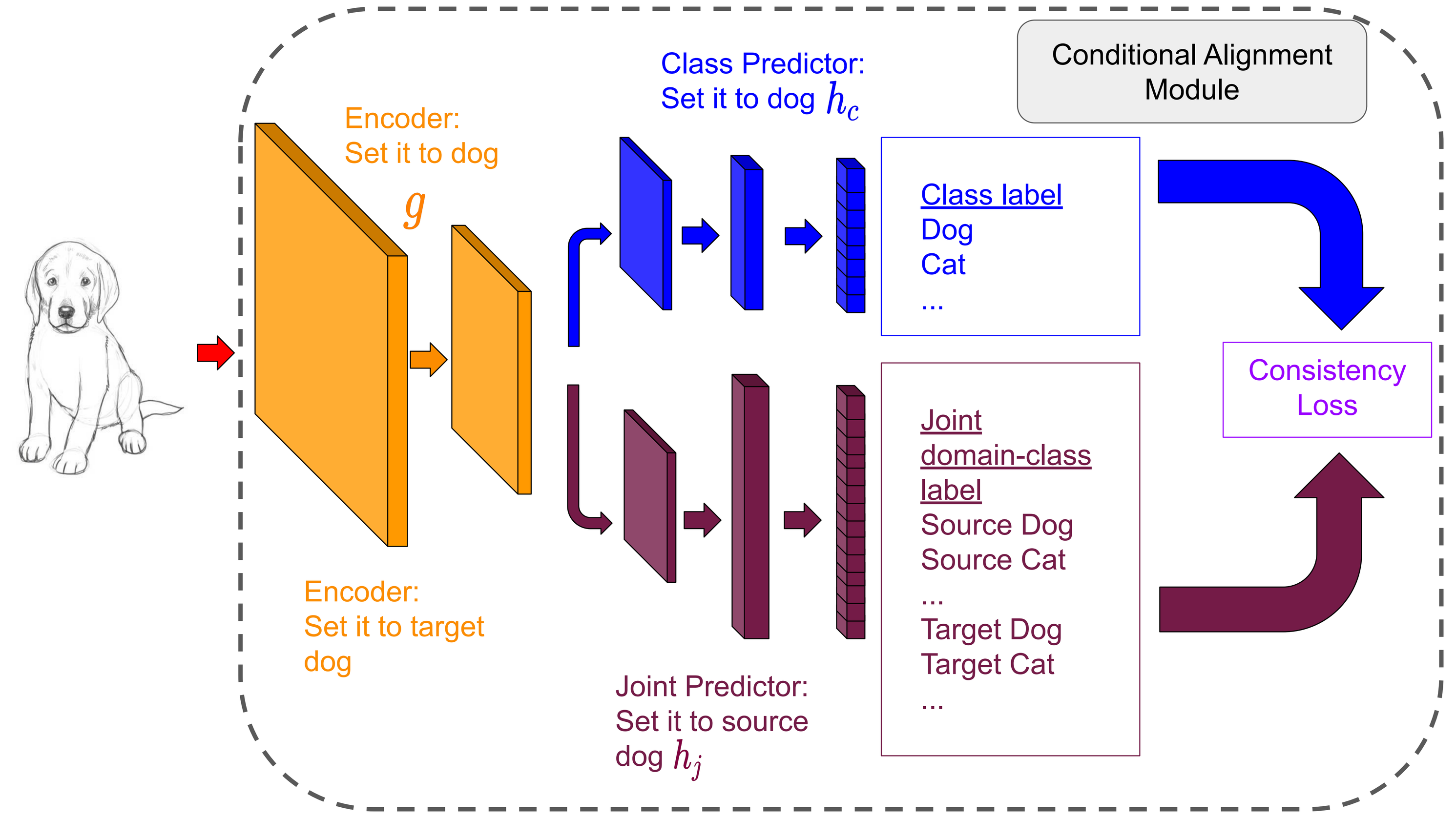

- Adversarial domain alignment (e.g. DANN) [1]

[1] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495, 2014.

## Conditional Alignment




## Proposed Method




## Objective Functions

- The source classification loss:

$$L_{sc}(f_c) = E_{(x,y)\sim P^s} \ell_{CE}(f_c(x), y)$$

- Pseudo-labels:

$$\hat{y} = e_k$$

$$k = \arg\max_k f_c(x)[k]$$

- The joint source and target classification losses:

$$L_{jsc}(h_j) = E_{(x,y)\sim P^s} \ell_{CE}(h_j(g(x)), [y,\ \mathbf{0}])$$

$$L_{jtc}(h_j) = E_{x\sim P^t_x} \ell_{CE}(h_j(g(x)), [\mathbf{0},\ \hat{y}])$$

- The joint source and target alignment losses:

$$L_{jsa}(g) = E_{(x,y)\sim P^s} \ell_{CE}(h_j(g(x)), [\mathbf{0},\ y])$$

$$L_{jta}(g) = E_{x\sim P^t_x} \ell_{CE}(h_j(g(x)), [\hat{y},\ \mathbf{0}])$$

The Joint Discriminator Feedback for Feature Alignment

## Theoretic Bounds for UDA [1]

$$\ell_t(h) \leq \ell_s(h) + \underbrace{\frac{1}{2}\hat{d}_{H\Delta H}(X^s, X^t)}_{1} + 4\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda$$

2

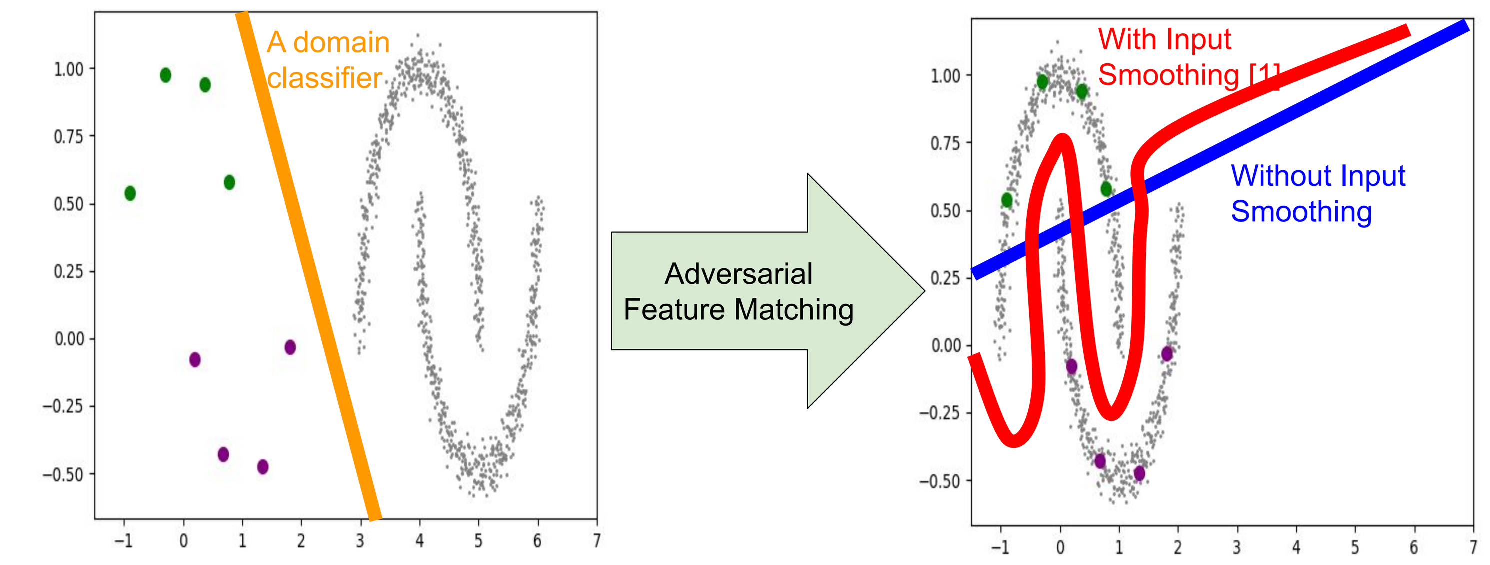
$$\lambda = \ell_s(h^*) + \ell_t(h^*)$$

$$h^* = \arg\min_{h\in\mathcal{H}} \ell_s(h) + \ell_t(h)$$

- DANN minimizes 1 but 2 can explode!
- With disjoint conditional alignment, our method also takes care of 2

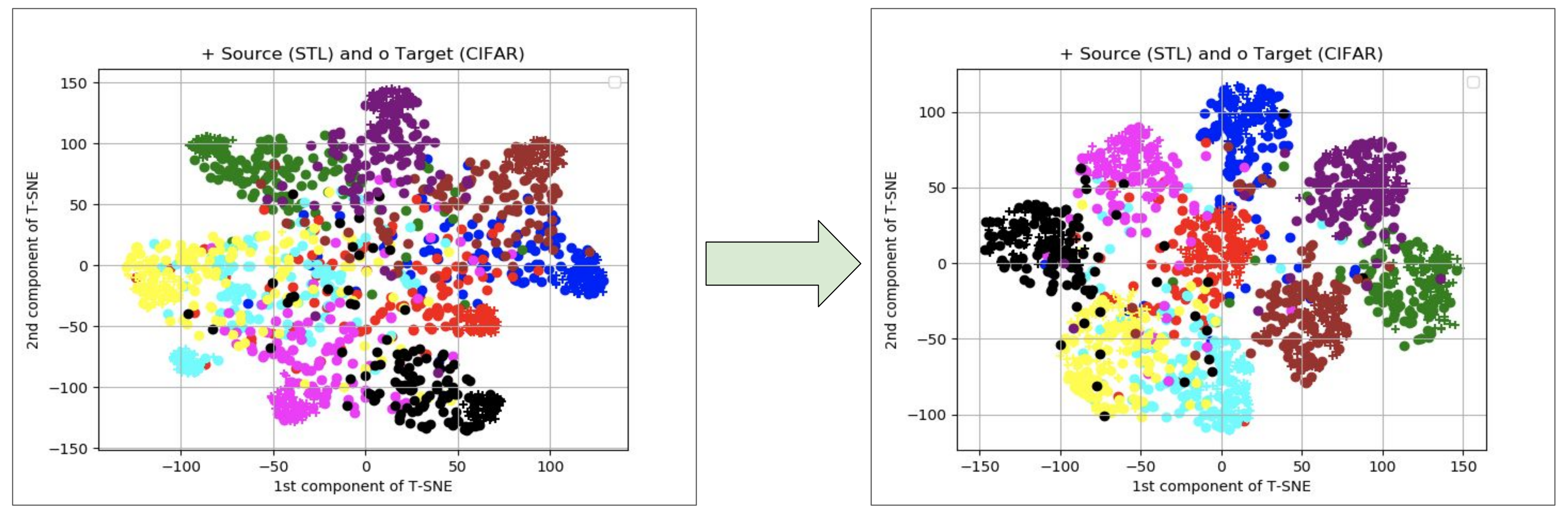
[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. Machine learning, 79(1-2):151–175, 2010.

## Exploiting Unlabeled Data with SSL Regularizers




[1] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. arXiv preprint arXiv:1704.03976, 2017.

## T-SNE [1]




[1] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.

## Analysis

**Proposition 1.** *The optimal joint predictor $h_j$ minimizing $L_{jsc}(h_j) + L_{jtc}(h_j)$ for any feature $z$ with non-zero measure either on $g\#P^s_x(z)$ or $g\#P^t_x(z)$ is*

$$h_j(z)[i] = \frac{g\#P^s(z, y=e_i)}{g\#P^s_x(z) + g\#P^t_x(z)}$$

$$h_j(z)[i+K] = \frac{g\#P^t(z, y=e_i)}{g\#P^s_x(z) + g\#P^t_x(z)} \text{ for } i \in \{1, ..., K\}$$

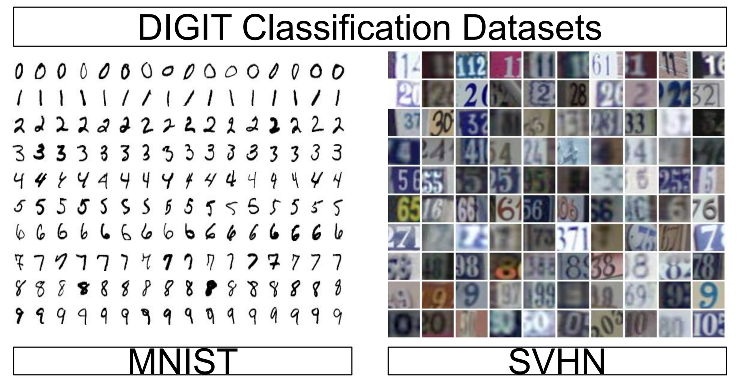
**Theorem 1.** *The objective $L_{jsa}(g) + L_{jta}(g)$ is minimized for the given optimal joint predictor if and only if*

$$g\#P^s(z|y=e_k) = g\#P^t(z|y=e_k)$$

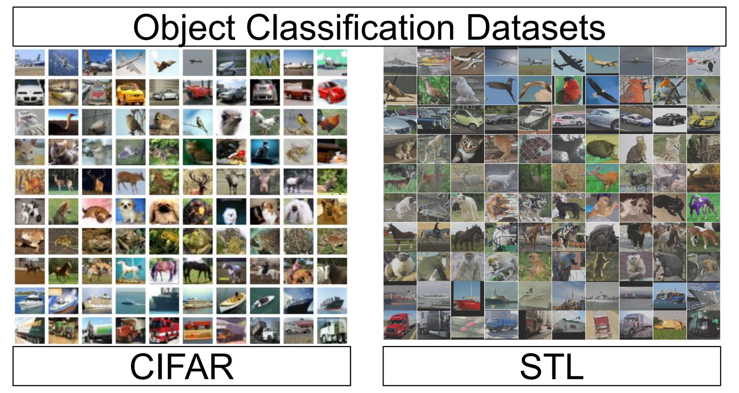
$$g\#P^s(z|y=e_k) > 0 \Rightarrow g\#P^s(z|y=e_i) = 0 \quad \text{for } i \neq k \text{ for any } y = e_k \text{ and } z.$$

## Datasets

DIGIT Classification Datasets




Object Classification Datasets




| Dataset | Number of training samples | Number of test samples | Resolution |
|---|---|---|---|
| MNIST | 60,000 | 10,000 | 28 by 28 |
| SVHN | 73,257 | 26,032 | 32 by 32 |
| CIFAR10 | 50,000 | 10,000 | 32 by 32 |
| STL | 5,000 | 8,000 | 96 by 96 |
| SYN-DIGITS | 479,400 | 9,553 | 32 by 32 |

## Comparison to SOA UDA Methods

| Source dataset | MNIST | SVHN | CIFAR | STL | SYN-DIGITS | MNIST |
|---|---|---|---|---|---|---|
| Target dataset | SVHN | MNIST | STL | CIFAR | SVHN | MNIST-M |
| DANN | 60.6 | 68.3 | 78.1 | 62.7 | 90.1 | 94.6 |
| VADA + IN [1] | 73.3 | 94.5 | 78.3 | 71.4 | 94.9 | 95.7 |
| DIRT-T +IN [1] | 76.5 | 99.4 | NR | a73.3 | 96.2 | 98.7 |
| Co-DA [2] | 81.7 | 99.0 | 81.4 | 76.4 | 96.4 | 99.0 |
| Co-DA + DIRT-T | 88.0 | **99.4** | NR | 77.6 | **96.4** | 99.1 |
| Ours | **89.19** | 99.33 | **81.65** | **77.76** | 96.22 | **99.47** |
| Source-only | 44.21 | 70.58 | 79.41 | 65.44 | 85.83 | 70.28 |
| Target-only | 94.82 | 99.28 | 77.02 | 92.04 | 96.56 | 99.87 |

[1] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. arXiv preprint arXiv:1802.08735, 2018.
[2] Abhishek Kumar, Prasanna Sattigeri, Kahini Wadhawan, Leonid Karlinsky, Rogerio Feris, Bill Freeman, and Gregory Wornell. Co-regularized alignment for unsupervised domain adaptation. In Advances in Neural Information Processing Systems, pages 9366–9377, 2018.