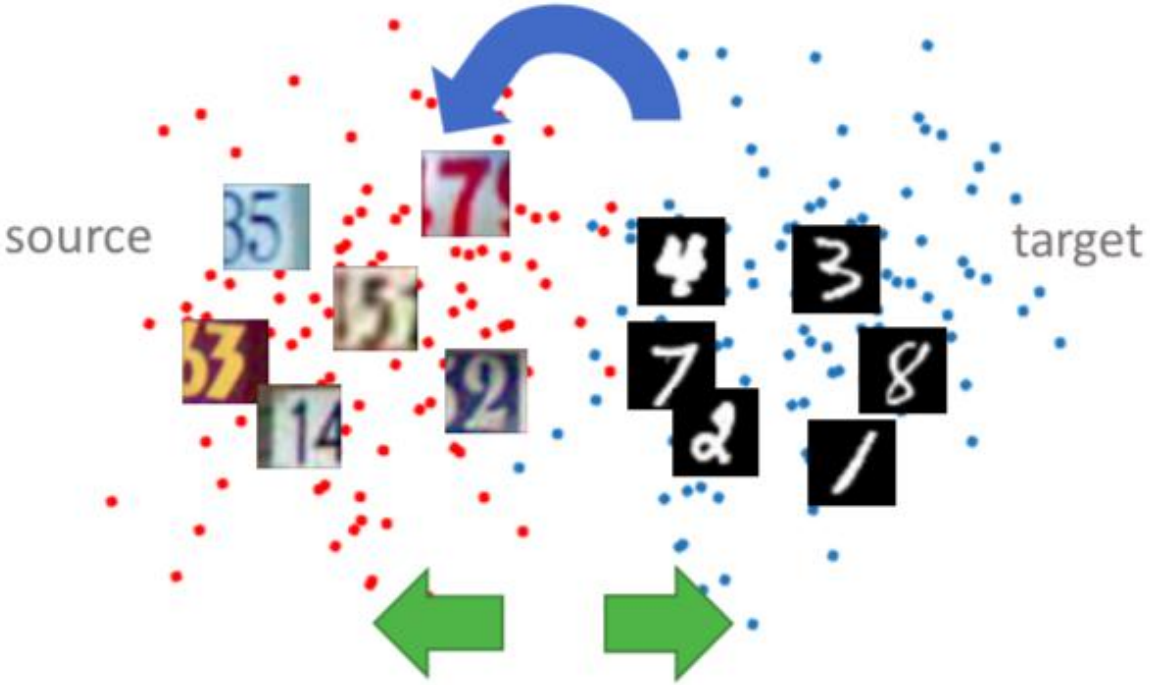
Cluster Alignment with a Teacher for Unsupervised Domain Adaptation

Zhijie Deng, Yucen Luo and Jun Zhu

01.03.2019

Unsupervised domain adaptation



Train on source domain: $\{x_i^s, y_i^s\}_{i=1,\ldots,N}$

Test on target domain: $\{x_i^t\}_{i=1,\ldots,M}$

Goal: classifier conquers the domain shift


Webcam

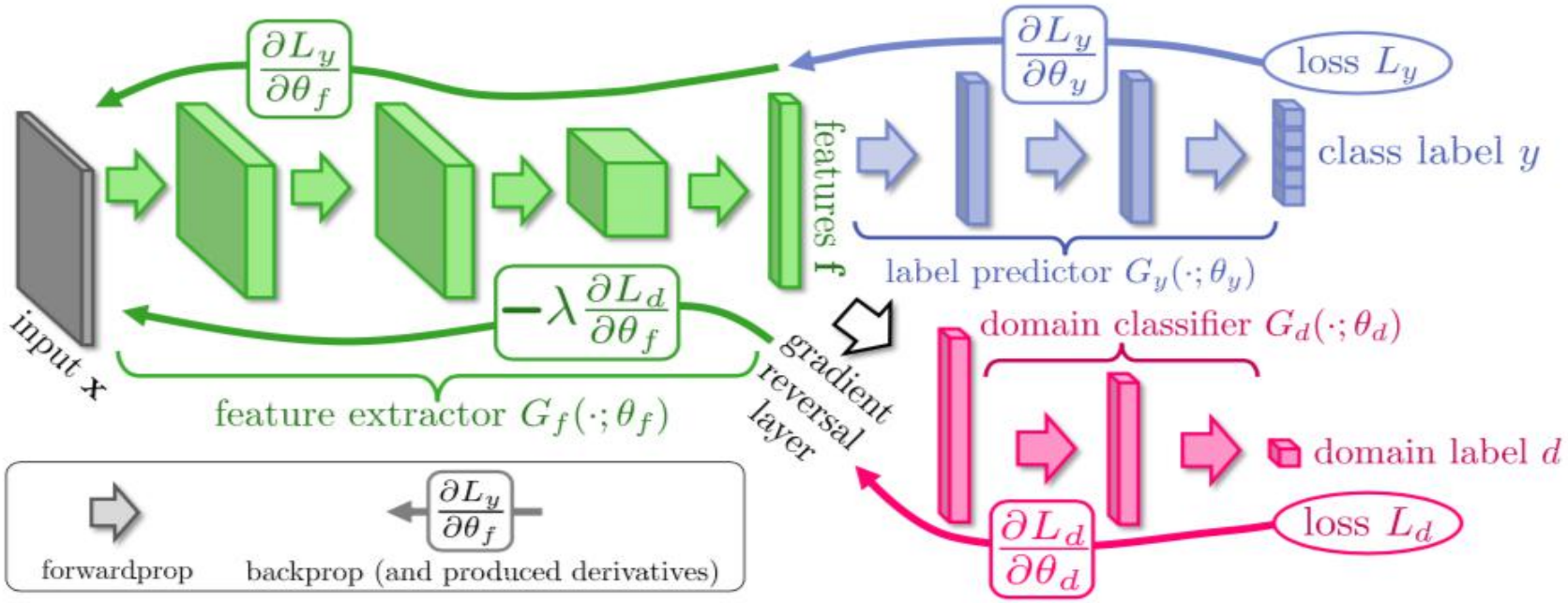

DSLR


Amazon

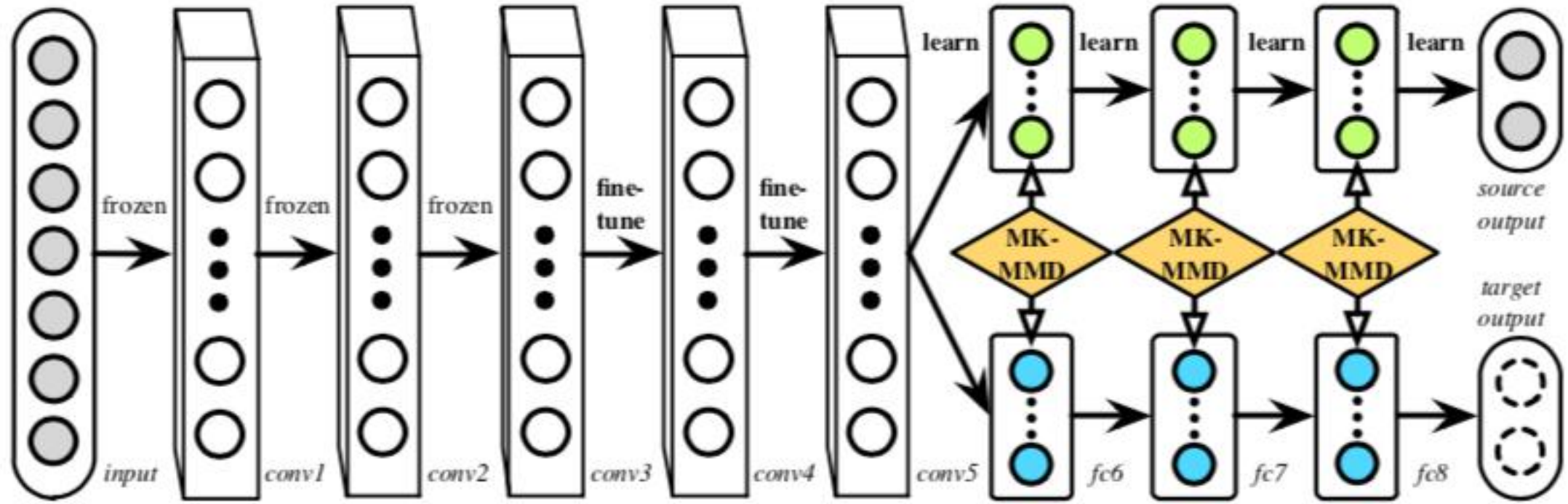

Caltech

Related work: domain adversarial training(Ganin and Lempitsky; Tzeng et al.)

- Based on Ben. David's the $\varepsilon_T(h) \leq \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(S, T) + C$.

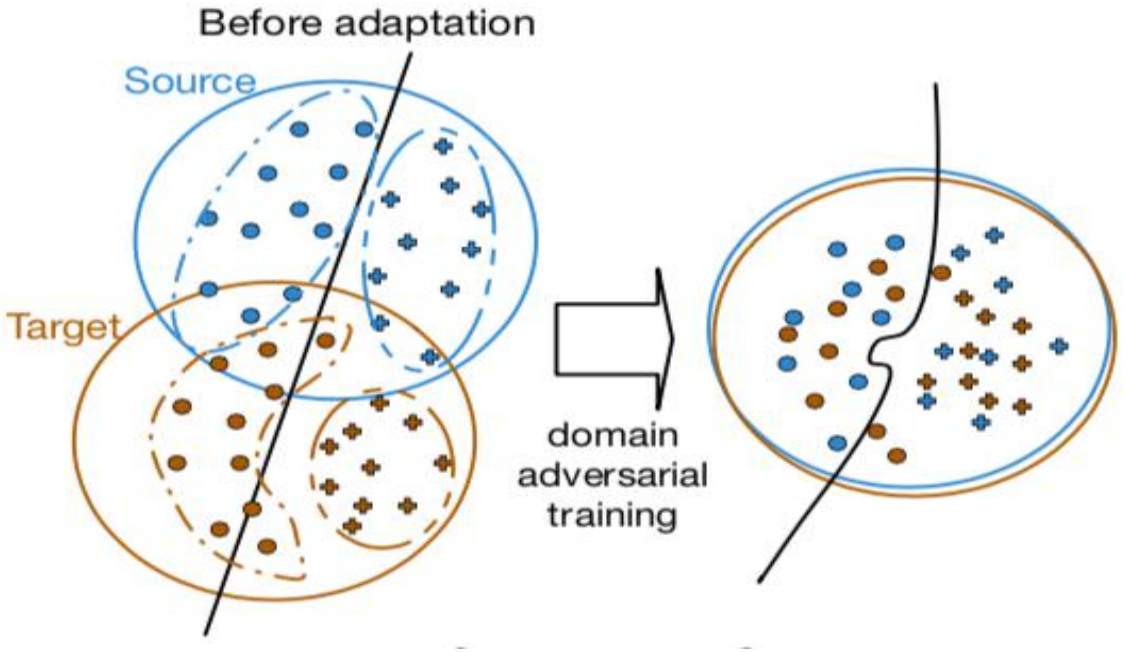


Related work: maximum mean discrepancy(MMD) based approaches(Long et al.)



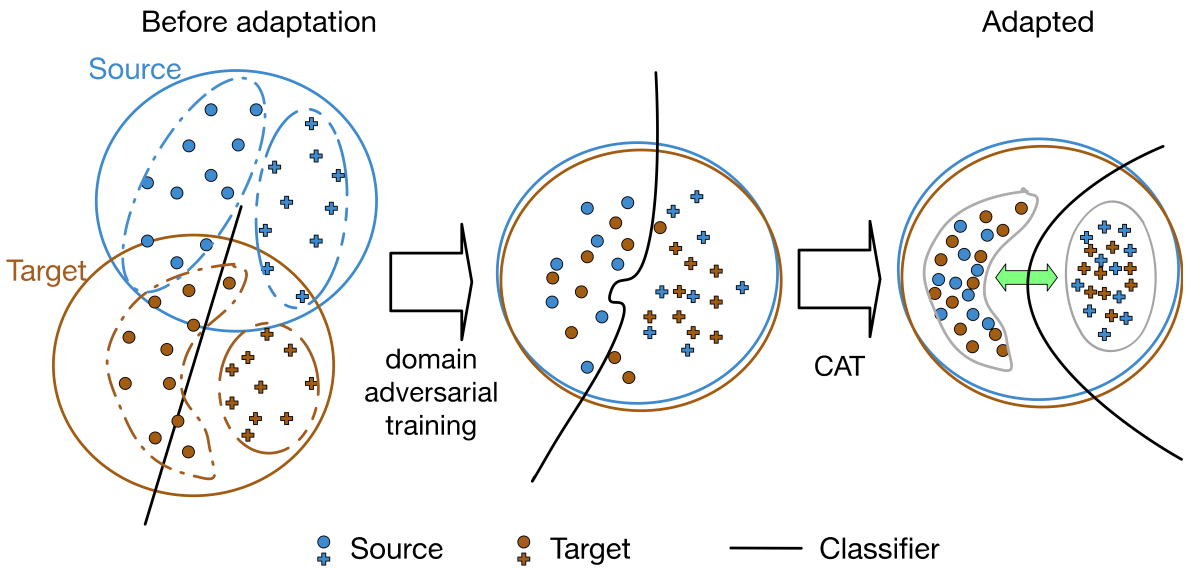
Issues of previous works

- Ignore the structure information of data manifolds in both the alignment and the classification process
- Lead to improper alignment and misclassification of target points



Motivation of CAT

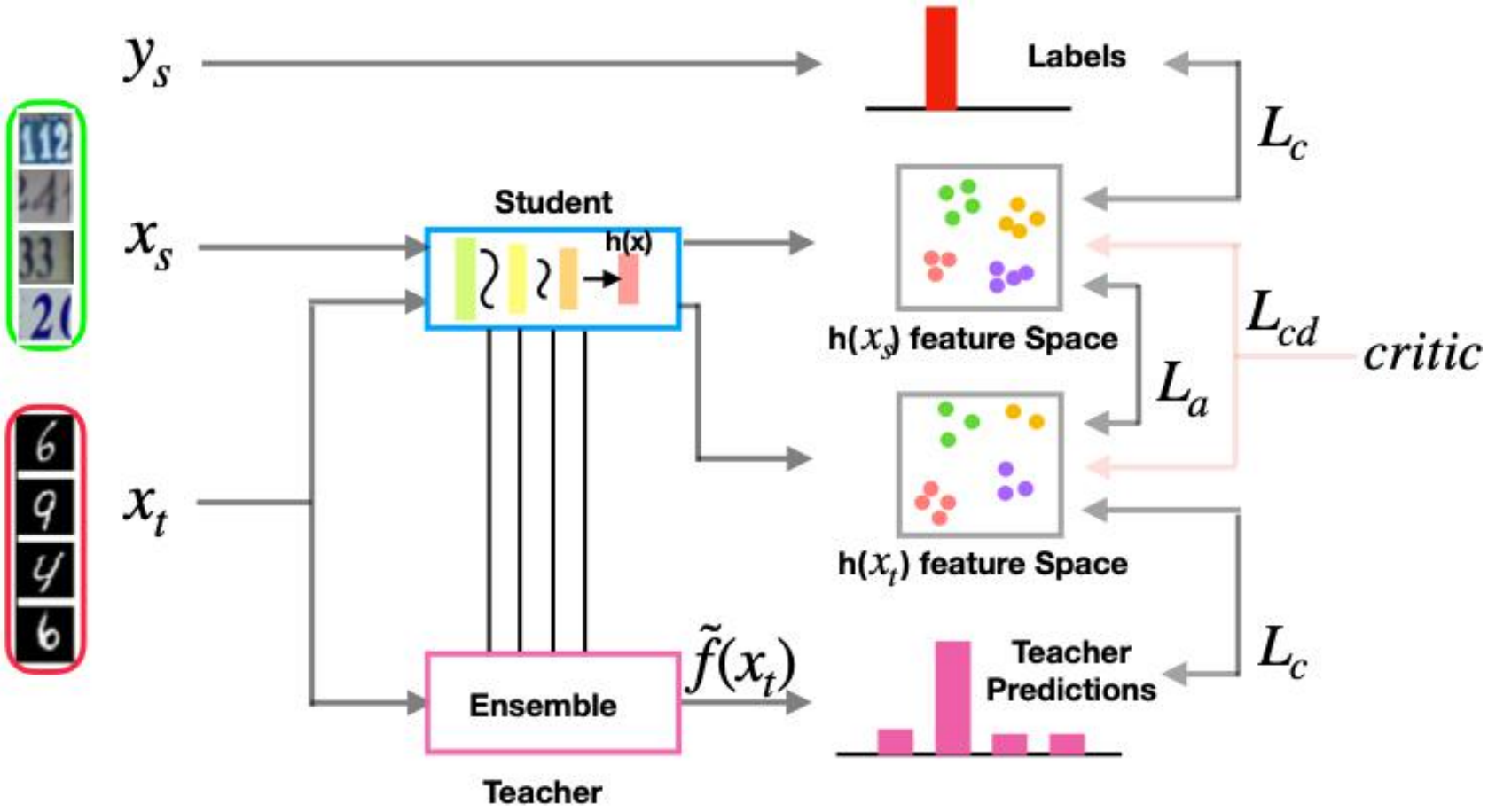
- Consider the cluster structures of the data manifolds
  - in the classification process: help to learn a feature space with improved discriminative power
  - in the alignment process: make the two domains aligned more properly and simplify the alignment problem between multi-mode domains



Why teacher?

- Classification based discriminative clustering
- Self evolution with one classifier
  - Sensitive to outliers and noise
  - A few wrongly classified instances may deteriorate the learning performance
- Clustering with a teacher classifier
  - Teacher generates different decision boundaries and has different abilities to learn
  - The error of student classifier will not be transferred back itself

Cluster Alignment with a Teacher (CAT)



Discover discriminative clusters in both domains: L_c

- Introduce a teacher model to predict the cluster alignments for target samples
- Use ground truth labels of source samples as their cluster alignments

$$\min_{\theta} \mathcal{L}_c(\mathcal{X}_s, \mathcal{X}_t) = \mathcal{L}_c(\mathcal{X}_s) + \mathcal{L}_c(\mathcal{X}_t),$$

- $$\mathcal{L}_c(\mathcal{X}) = \frac{1}{|\mathcal{X}|^2} \sum_{i=1}^{|\mathcal{X}|} \sum_{j=1}^{|\mathcal{X}|} \left[\delta_{ij} d\left(h(x^i), h(x^j)\right) + (1 - \delta_{ij}) \max\left(0, m - d\left(h(x^i), h(x^j)\right)\right) \right],$$

- Encourages the features from the same cluster to concentrate together and pushes the features from different clusters far away from each other with a distance m at least

Cluster alignment with centers: L_a

- Align the cluster-structure distributions through matching the

$$\min_{\theta} \mathcal{L}_a(\mathcal{X}_s, \mathcal{Y}_s, \mathcal{X}_t) = \frac{1}{K} \sum_{k=1}^{K} [d(\lambda_{s,k}, \lambda_{t,k})]$$

- $$\lambda_{s,k} = \frac{1}{|\mathcal{X}_{s,k}|} \sum_{x_s^i \in \mathcal{X}_{s,k}} h(x_s^i),\ \lambda_{t,k} = \frac{1}{|\mathcal{X}_{t,k}|} \sum_{x_t^i \in \mathcal{X}_{t,k}} h(x_t^i)$$

- The features from the same class but different domains tend to concentrate together
- Harmful signals of incorrect labeled target samples can be alleviated by the summation operation and the robust teacher

Reduce uncertainty in adversarial training: L_{cd}

- In the early stages of training, quite a number of target samples lie around the decision boundaries and have high probability to be mapped into wrong clusters
- we propose a confidence-based variant of the domain adversarial loss:

$$\min_{\theta}\max_{\phi}\mathcal{L}_{cd}(\mathcal{X}_s,\mathcal{X}_t)=\frac{1}{N}\sum_{i=1}^{N}\left[\log c\left(h(x_s^i;\theta);\phi\right)\right]+$$
$$\frac{1}{\tilde{M}}\sum_{i=1}^{\tilde{M}}\left[\log\left(1-c\left(h(x_t^i;\theta);\phi\right)\right)\gamma_i\right]$$

- This improves the correctness of learned clusters and enhances the stability of training, bringing better generalization performance of the classifier in the target domain

Overall objective

$$\min_{\theta} \max_{\phi} \mathcal{L}_y + \alpha \mathcal{L}_{cd} + \beta(\mathcal{L}_c + \mathcal{L}_a)$$

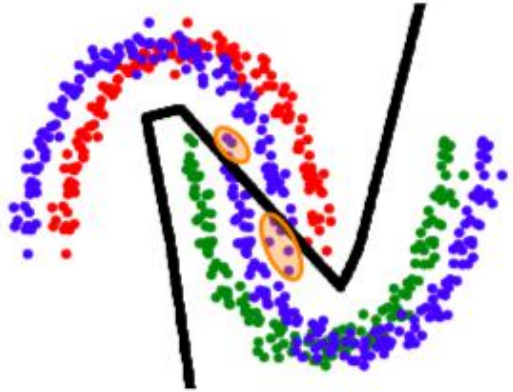
Insights from theoretical analysis

$$\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(S,T) + C, \quad \forall h \in \mathcal{H}$$

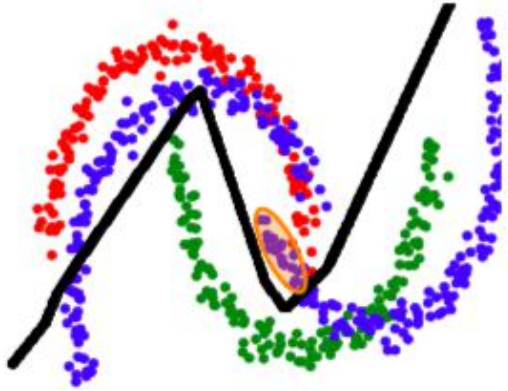
$$C \leq \min_{h \in \mathcal{H}} \epsilon_{\mathcal{S}}(h, f_{\mathcal{S}}) + \epsilon_{\mathcal{T}}(h, f_{\mathcal{S}}) + \epsilon_{\mathcal{T}}(f_{\mathcal{S}}, f_{\hat{\mathcal{T}}}) + \epsilon_{\mathcal{T}}(f_{\mathcal{T}}, f_{\hat{\mathcal{T}}})$$

- Minimizing the supervised loss in source domain can reduce the first two terms effectively
- The third term represents the inconsistency between f_S and $f_{T^\wedge}$ which is minimized by discovering discriminative cluster structures of domains and then reducing the distance between corresponding pairs
- The last term is minimized by CAT using the mutual boosting cycle between the student classifier and its teacher

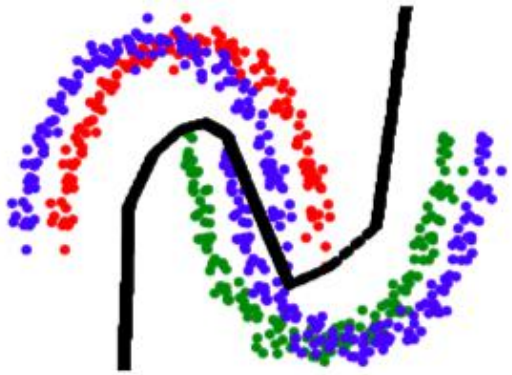
Experiments: synthetic data



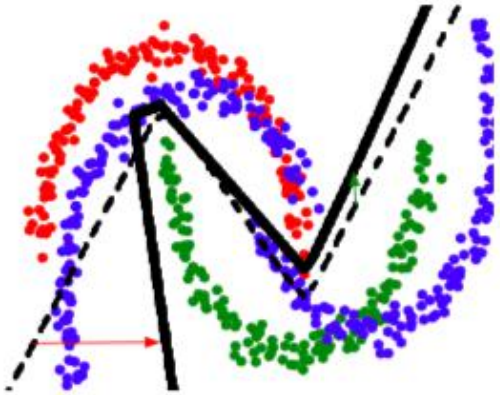
(a) RevGrad: failure case 1



(b) RevGrad: Failure Case 2



(c) CAT



(d) CAT

Experiments: digits datasets

Method	*SVHN* to *MNIST*	*MNIST* to *USPS*	*USPS* to *MNIST*
Source Only	60.1 ± 1.1	75.2 ± 1.6	57.1 ± 1.7
RevGrad [8]	73.9	77.1 ± 1.8	73.0 ± 2.0
DDC [39]	68.1 ± 0.3	79.1 ± 0.5	66.5 ± 3.3
CoGAN [20]	-	91.2 ± 0.8	89.1 ± 0.8
DRCN [9]	82.0 ± 0.1	91.8 ± 0.09	73.7 ± 0.04
ADDA [38]	76.0 ± 1.8	89.4 ± 0.2	90.1 ± 0.8
LEL [26]	81.0 ± 0.3	-	-
MCD [34]	96.2 ± 0.4	94.2 ± 0.7	94.1 ± 0.3
MSTN [42]	91.7 ± 1.5	92.9 ± 1.1	-
CAT(ours)	$\mathbf{98.4} \pm 0.6$	$\mathbf{95.0} \pm 0.8$	$\mathbf{96.5} \pm 0.5$

Experiments: office-31 and ImageCLEF-D^

Method	A to W	D to W	W to D	A to D	D to A	W to A	Avg
AlexNet [15]	61.6 ± 0.5	95.4 ± 0.3	99.0 ± 0.2	63.8 ± 0.5	51.1 ± 0.6	49.8 ± 0.4	70.1
DDC [39]	61.8 ± 0.4	95.0 ± 0.5	98.5 ± 0.4	64.4 ± 0.3	52.1 ± 0.6	52.2 ± 0.4	70.6
DRCN [9]	68.7 ± 0.3	96.4 ± 0.3	99.0 ± 0.2	66.8 ± 0.5	56.0 ± 0.5	54.9 ± 0.5	73.6
RevGrad [8]	73.0 ± 0.5	96.4 ± 0.3	99.2 ± 0.3	72.3 ± 0.3	53.4 ± 0.4	51.2 ± 0.5	74.3
RTN [24]	73.3 ± 0.3	96.8 ± 0.2	99.6 ± 0.1	71.0 ± 0.2	50.5 ± 0.3	51.0 ± 0.1	73.7
JAN [23]	74.9 ± 0.3	96.6 ± 0.2	99.5 ± 0.2	71.8 ± 0.2	58.3 ± 0.3	55.0 ± 0.4	76.0
AutoDIAL [3]	75.5	96.6	99.5	73.6	58.1	59.4	77.1
MSTN [42]	80.5 ± 0.4	96.9 ± 0.1	99.9 ± 0.1	74.5 ± 0.4	62.5 ± 0.4	60.0 ± 0.6	79.1
CAT (ours)	**80.7** ± 1.6	**97.6** ± 0.1	**100.0** ± 0.0	**76.4** ± 0.6	**63.7** ± 0.5	**62.2** ± 0.4	**80.1**

Table 1: Summary of domain adaptation results on the *Office-31* datasets in terms of test accuracy (%). (AlexNet)

Method	I to P	P to I	I to C	C to I	C to P	P to C	Avg
AlexNet [15]	66.2 ± 0.2	70.0 ± 0.2	84.3 ± 0.2	71.3 ± 0.4	59.3 ± 0.5	84.5 ± 0.3	73.9
RevGrad [8]	66.5 ± 0.5	81.8 ± 0.4	89.0 ± 0.5	79.8 ± 0.5	63.5 ± 0.4	88.7 ± 0.4	78.2
RTN [24]	67.4 ± 0.3	81.3 ± 0.3	89.5 ± 0.4	78.0 ± 0.2	62.0 ± 0.2	89.1 ± 0.1	77.9
JAN [23]	67.2 ± 0.5	82.8 ± 0.4	91.3 ± 0.5	80.0 ± 0.5	63.5 ± 0.4	91.0 ± 0.4	79.3
MSTN [42]	67.3 ± 0.3	82.8 ± 0.2	91.5 ± 0.1	**81.7** ± 0.3	65.3 ± 0.2	91.2 ± 0.2	80.0
CAT (ours)	**68.6** ± 0.1	**84.6** ± 0.5	**91.9** ± 0.4	80.8 ± 0.3	**65.6** ± 0.6	**92.5** ± 0.2	**80.7**

Table 2: Summary of domain adaptation results on the *ImageCLEF-DA* datasets in terms of test accuracy (%). (AlexNet)

Experiments: office-31 and ImageCLEF-DA

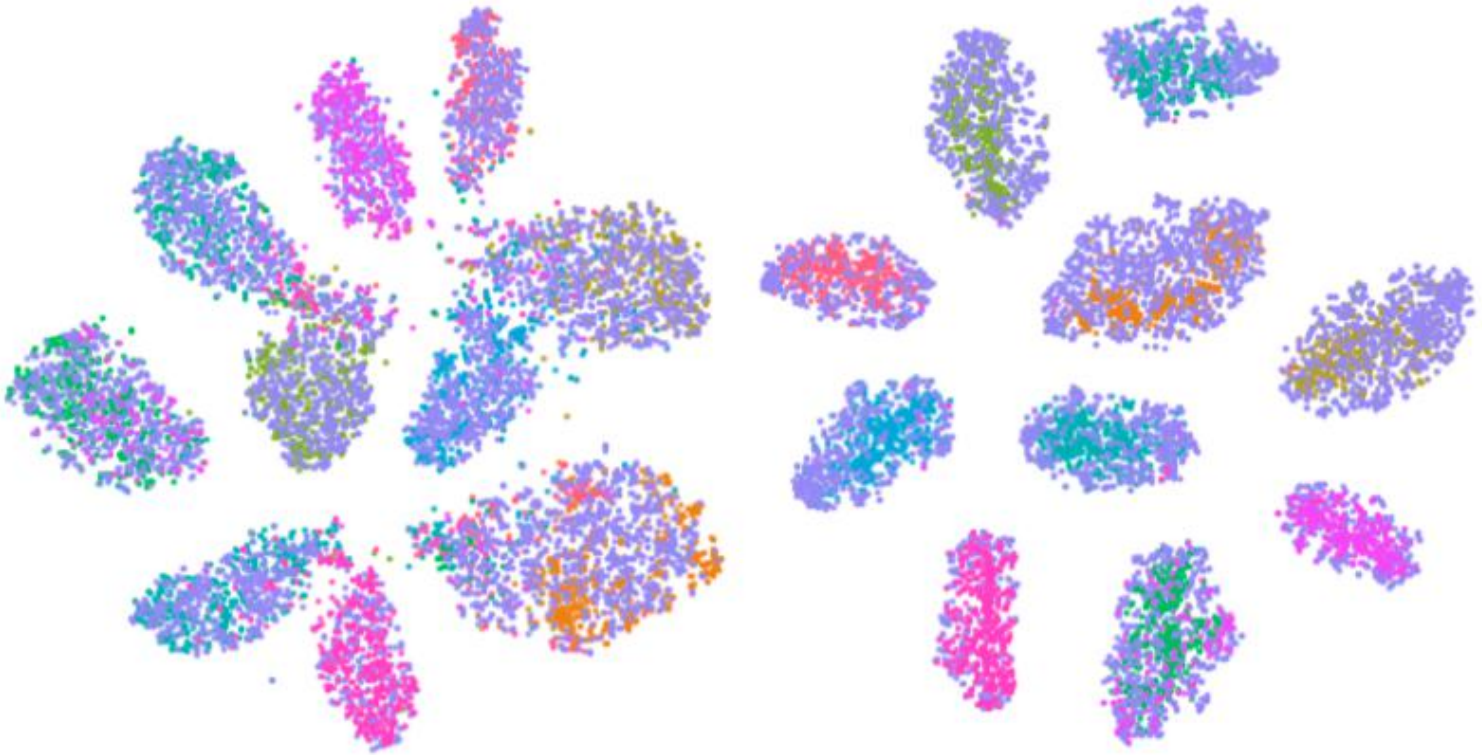
Method	A to W	D to W	W to D	A to D	D to A	W to A	Avg
ResNet-50 [2]	68.4 ± 0.2	96.7 ± 0.1	99.3 ± 0.1	68.9 ± 0.2	62.5 ± 0.3	60.7 ± 0.3	76.1
DAN [4]	80.5 ± 0.4	97.1 ± 0.2	99.6 ± 0.1	78.6 ± 0.2	63.6 ± 0.3	62.8 ± 0.2	80.4
RevGrad [1]	82.0 ± 0.4	96.9 ± 0.2	99.1 ± 0.1	79.4 ± 0.4	68.2 ± 0.4	67.4 ± 0.5	82.2
RTN [6]	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
GenToAdapt [7]	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
JAN [5]	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
Modified JAN (ours)	94.0 ± 0.4	96.6 ± 0.6	**100.0** ± 0.0	88.1 ± 1.0	68.9 ± 0.7	69.4 ± 0.5	86.2
CAT (ours)	**94.4** ± 0.1	**98.0** ± 0.2	**100.0** ± 0.0	**90.8** ± 1.8	72.2 ± 0.6	70.2 ± 0.1	**87.6**

Table 1: Summary of domain adaptation results on the *Office-31* datasets in terms of test accuracy (%). (ResNet-50)

Method	I to P	P to I	I to C	C to I	C to P	P to C	Avg
ResNet-50 [2]	74.8 ± 0.3	83.9 ± 0.1	91.5±0.3	78.0 ± 0.2	65.5 ± 0.3	91.2 ± 0.3	80.7
DAN [4]	74.5 ± 0.4	82.2 ± 0.2	92.8 ± 0.2	86.3 ± 0.4	69.2 ± 0.4	89.8 ± 0.4	82.5
RevGrad [1]	75.0 ± 0.6	86.0 ± 0.3	**96.2** ± 0.4	87.0 ± 0.5	74.3 ± 0.5	91.5 ± 0.6	85.0
JAN [5]	76.8 ± 0.4	88.0 ± 0.2	94.7 ± 0.2	89.5 ± 0.3	74.2 ± 0.3	91.7 ± 0.3	85.8
Modified JAN (ours)	76.3 ± 0.8	89.2 ± 0.8	95.3 ± 0.7	89.3 ± 0.3	**75.9** ± 1.1	92.2 ± 1.3	86.4
CAT (ours)	**77.2** ± 0.2	**91.0** ± 0.3	95.5 ± 0.3	**91.3** ± 0.3	75.3 ± 0.6	**93.6** ± 0.5	**87.3**

Table 2: Summary of domain adaptation results on the *ImageCLEF-DA* datasets in terms of test accuracy (%). (ResNet-50)

Experiments: Visualization of manifolds

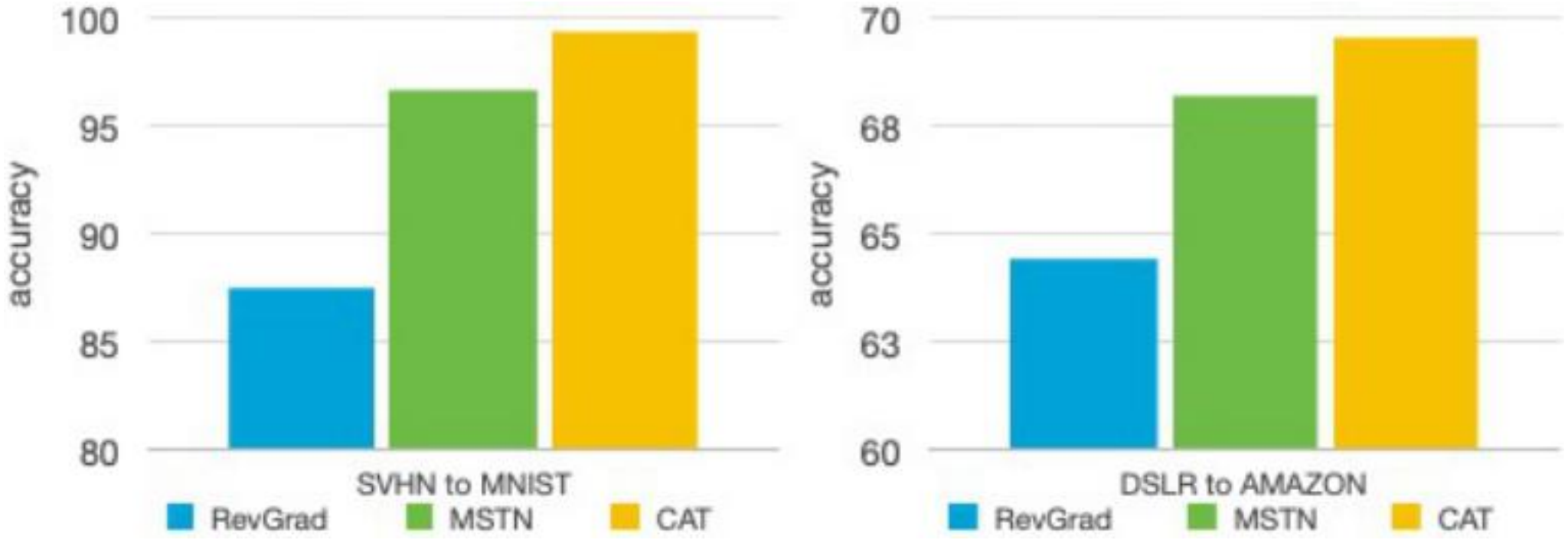


(a) RevGrad (b) CAT

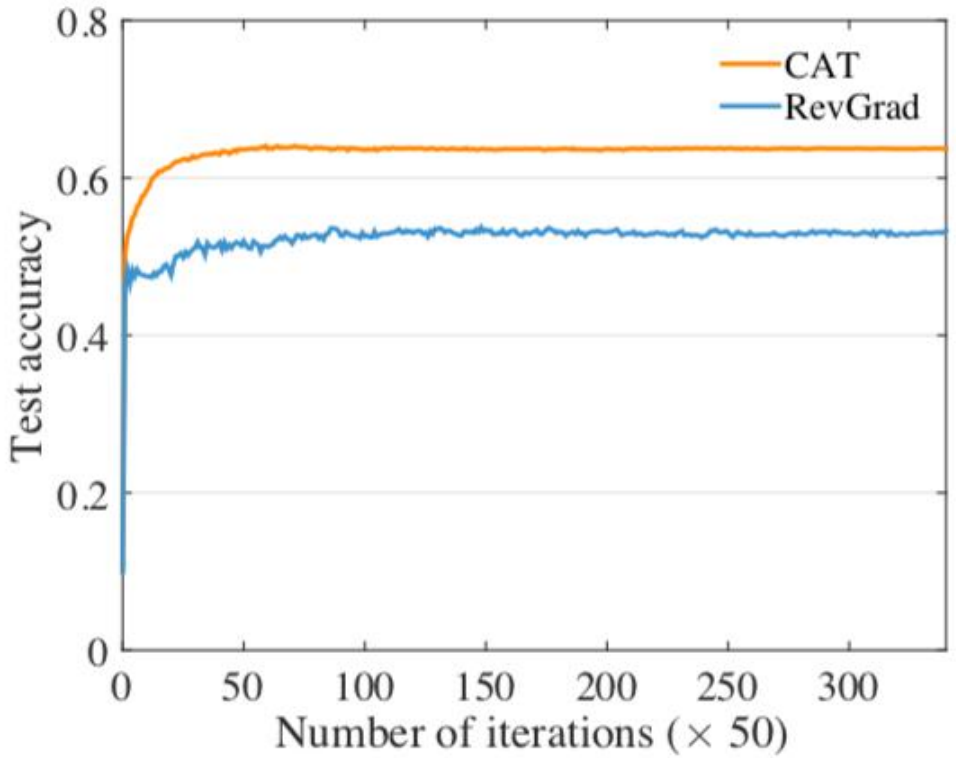
Experiments: ablation study of L_c L_a and L_{cd}

Method	*SVHN* to *MNIST*
RevGrad($\mathcal{L}_y + \alpha\mathcal{L}_d$)	73.9
$\mathcal{L}_y + \alpha\mathcal{L}_{cd}$	77.1 ± 1.2
$\mathcal{L}_y + \alpha\mathcal{L}_d + \beta\mathcal{L}_c$	97.6 ± 1.5
$\mathcal{L}_y + \alpha\mathcal{L}_{cd} + \beta\mathcal{L}_c$	97.8 ± 1.4
$\mathcal{L}_y + \alpha\mathcal{L}_d + \beta\mathcal{L}_a$	96.3 ± 0.8
$\mathcal{L}_y + \alpha\mathcal{L}_{cd} + \beta\mathcal{L}_a$	97.4 ± 0.2
CAT($\mathcal{L}_y + \alpha\mathcal{L}_{cd} + \beta(\mathcal{L}_c + \mathcal{L}_a)$)	$\mathbf{98.4} \pm 0.6$

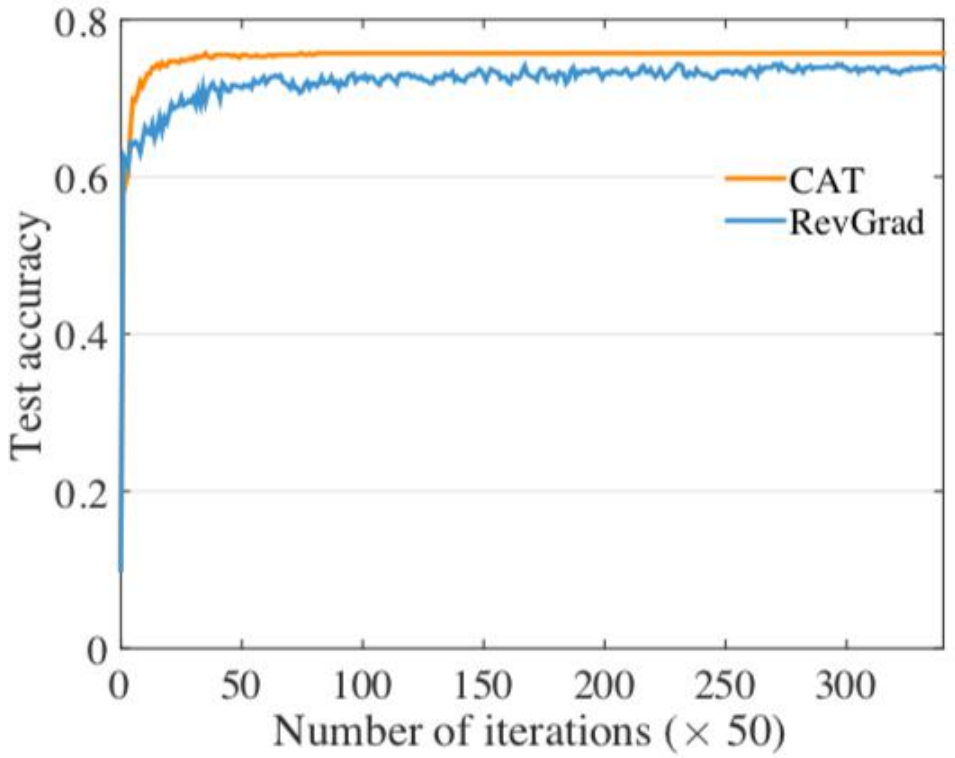
Experiments: Clustering in the feature space



Experiments: Convergence



(a) DSLR to AMAZON



(b) AMAZON to DSLR

Conclusion

- Cluster Alignment with a Teacher
  - discovers the underlying cluster structures of data manifolds
  - aligns source domain with target domain better based on it
  - produces a domain-invariant feature space with improved discriminative power
  - enhances the domain adaptation results significantly
  - establishes new state-of-the-art results on several standard benchmarks

Thanks!