Supplementary Material:
Cluster Alignment with a Teacher for Unsupervised Domain Adaptation

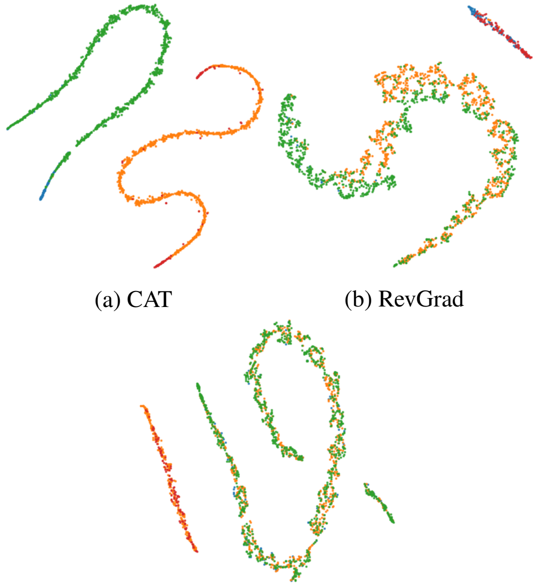
Zhijie Deng, Yucen Luo, Jun Zhu*
Dept. of Comp. Sci. & Tech., Institute for AI, BNRist Lab, THBI Lab, Tsinghua University
{dzj17, luoyc15}@mails.tsinghua.edu.cn, dcszj@tsinghua.edu.cn

A. Class-conditional cluster structure on more tasks

First, we visualize the learned feature spaces of CAT, RevGrad [7] and MSTN [49] on the imbalanced *SVHN* to *MNIST* task using t-SNE [27], as shown in Fig.1. It is obvious that CAT can force the samples from the same class to concentrate together to form tighter clusters than those of RevGrad and MSTN, and the clusters present strip pattern in the 2-D space. CAT can also align the class-conditional distributions of the source and the target domains correctly. However, RevGrad and MSTN tend to align the '0' images in *SVHN* with the '1' images in *MNIST*, thus the learned feature spaces of them are confusing and not discriminative. This visualization verifies the results in Table. 1.

Second, we plot the feature spaces learned by CAT+rRevGrad and RevGrad on *MNIST* to *USPS* and *USPS* to *MNIST* tasks in Fig. 2 using t-SNE [27]. CAT+rRevGrad can deliver more discriminative feature spaces with separable and tight class-conditional clusters. Therefore, it is sufficient to use the first-order statistics based matching loss $\mathcal{L}_a$ to match the class-conditional distributions of the two domains. The aligned clusters of the source and the target domains also verify the effectiveness of the loss $\mathcal{L}_a$.

Furthermore, we examine the feature space learned by CAT on more challenging tasks in *Office-31* dataset and *ImageCLEF-DA* dataset, and results are demonstrated in Fig. 3. These features are outputs of AlexNet trained with rRevGrad+CAT. The class-conditional distributions are shaped to be tight and separable clusters, and the corresponding cluters from the source domain and the target domain are aligned. Therefore, CAT can achieve the objectives of discriminative learning and class-conditional alignment, thus can perform well on the extensive experiments on *Office-31* and *ImageCLEF-DA* datasets.



(a) CAT

(b) RevGrad

(c) MSTN

Figure 1: (Best viewed in color.) Feature space learned on imbalanced *SVHN* to *MNIST* task. Green, red, blue and orange points represent '0' images from *SVHN*, '1' images from *SVHN*, '0' images from *MNIST* and '1' images from *MNIST*, respectively.

B. Quantitative estimate of the divergence between domains

When aligning the source domain and target domain via the combination of RevGrad and CAT, the loss $\mathcal{L}_d$ which is maximized w.r.t. the critic c can be viewed as a lower bound of $2JSD(s,t) - 2\log 2$ (see [9] for the details) where JSD denotes the Jensen-Shannon divergence between distributions. Therefore, we plot $\frac{1}{2}\mathcal{L}_d + \log 2$ to quantitatively estimate the divergence between the two domains, following [49]. The results are shown in Fig. 4 and we use the AlexNet as the classifier here. CAT can boost RevGrad significantly, leading to faster and better convergence. This group of experiments verifies that when combining CAT with the marginal distribution alignment approaches, it can provide a

*Corresponding author.

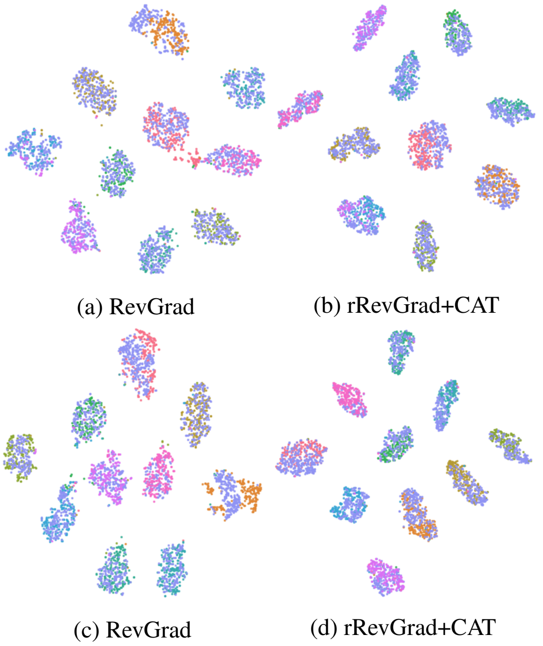
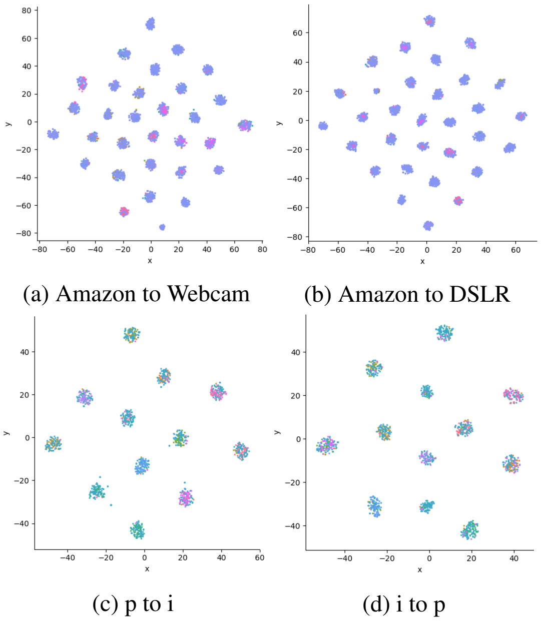

(a) RevGrad
(b) rRevGrad+CAT
(c) RevGrad
(d) rRevGrad+CAT

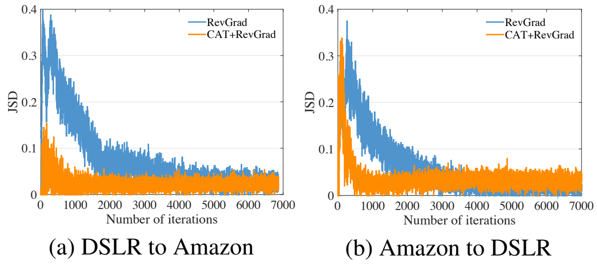
Figure 2: (Best viewed in color.) Feature space learned on *MNIST* to *USPS* (Fig. 2a and Fig. 2b) and *USPS* to *MNIST* (Fig. 2c and Fig. 2d) tasks. Blue violet denotes the source domain and the other colors denote different classes of target domain.



(a) Amazon to Webcam
(b) Amazon to DSLR
(c) p to i
(d) i to p

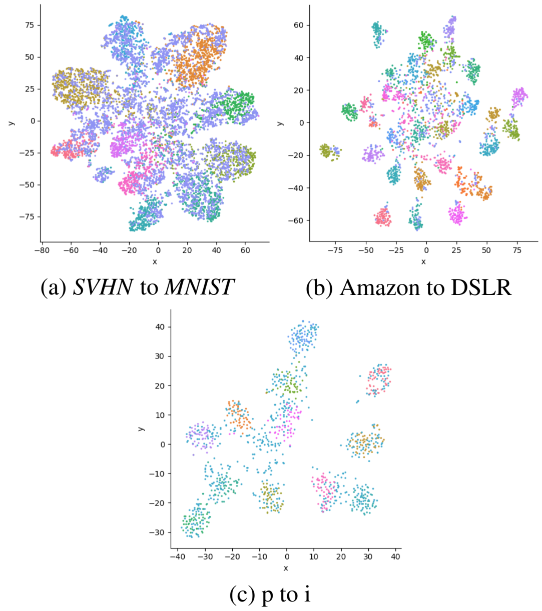
Figure 3: (Best viewed in color.) Feature space learned on four challenging tasks. Blue violet (in (a) and (b)) and deep sky blue (in (c) and (d)) denote the source domain and the other colors denote different classes of target domain.

discriminative class-conditional alignment and bias the existing approaches to align the cluster-structure marginal distributions better.



(a) DSLR to Amazon
(b) Amazon to DSLR

Figure 4: Jensen-Shannon divergence (JSD) curves during training.

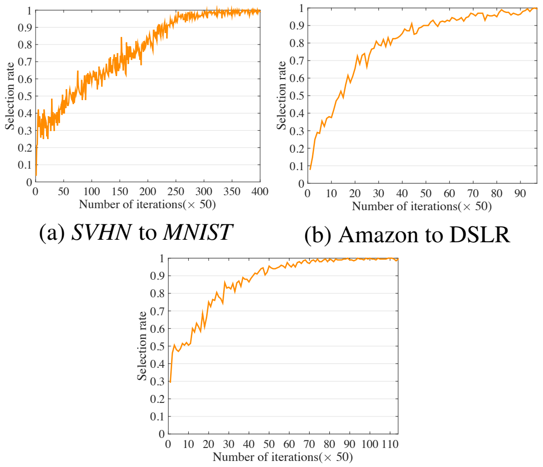


(a) *SVHN* to *MNIST*
(b) Amazon to DSLR
(c) p to i

Figure 5: (Best viewed in color.) Feature space in the early stages of training. **Different** from the above feature spaces, blue violet (in (a) and (b)) and deep sky blue (in (c)) denote the **target** domain and the other colors denote different classes of **source** domain.

C. Verification of confidence-thresholding technique

Since the source classification loss and the source discriminative clustering loss can produce strong gradients and converge quickly, the discriminative cluster structure will form in the source domain in the early stages of training. However, the classifier has not been adapted for the target domain, so a notable part of the target features will lie in the gaps between the source clusters and have low classification confidence. Therefore, the marginal alignment approaches may easily map these features into incorrect clusters, as stated in Sec. 3.2.3. To address this problem, we propose the confidence-thresholding technique which includes the fine-level structure information into marginal alignment approaches. We claim that in the training pro-



(a) *SVHN* to *MNIST*

(b) Amazon to DSLR

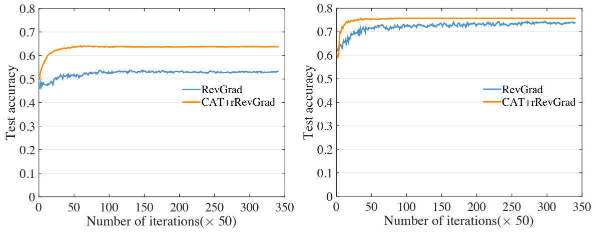
(c) p to i

Figure 6: The selection rate of the confidence-thresholding technique on different tasks.

cedure, the discriminative class-conditional alignment between the two domains forms gradually, so more and more samples are going to be selected into the marginal alignment training. Here we prove these through experiments on tasks in *SVHN-MNIST-USPS*, *Office-31* and *ImageCLEF-DA*. At first, we train the RevGrad+CAT models on the three tasks with limited iterations (*i.e.*, 2000 iterations on *SVHN* to *MNIST* task, 100 iterations on Amazon to DSLR task and 100 iterations on p to i task) and plot the feature spaces of them in Fig 5. Obviously, a notable part of target samples lie in the gaps between the source clusters, especially on the *SVHN* to *MNIST* task which has large source and target domains. Then, we train the rRevGrad+CAT models on these tasks following the same settings, and we plot the selection rate of the confidence-thresholding technique w.r.t. the number of iterations in Fig. 6. When using this technique, we note that the selection rate monotonically increases with the number of iterations and after several thousands of iterations, the selection rate will be almost 100% on the Amazon to DSLR and p to i tasks. On *SVHN* to *MNIST* task, we use a ramp-up function $exp(-10 * (1 - min(\frac{ite-5000}{10000}, 1.)))$ as α after 5000 iterations, suggested by related SSL works. Therefore, after around 15000 iterations, the discriminative clustering structure forms, and then the samples are pushed far away from the decision boundaries. So almost all the samples will have confidence more than p and will be selected into the domain adversarial training.

D. Convergence

To inspect how CAT converges, we plot the test accuracy with respect to the number of iterations in Fig. 7. On the two adaptation tasks using AlexNet, CAT shows similar convergence rate with RevGrad [7] but better performance.



(a) DSLR to Amazon

(b) Amazon to DSLR

Figure 7: Test accuracy curves.

E. Experimental details

On digits adaptation tasks, we use the simple LeNet with Batch Normalization after the convolutional layers and use the probability logits as features for adaptation, following [49, 44]. When combining with RevGrad [7] and rRevGrad, the critic model has a $10 \to 500 \to 500 \to 1$ architecture.

On more challenging tasks, we conduct experiments based on the AlexNet [15] and ResNet-50 [12] equipped with 256-D bottleneck layers after the $fc7$ and $pool5$ layers respectively (following [24, 49]). We use the features outputted by the bottleneck layers as image representations for adaptation and use a three-layer critic with $256 \to 1024 \to 1024 \to 1$ architecture. We finetune all the layers before the bottleneck layers in AlexNet and ResNet-50 and train the bottleneck layers and the classification layers via back propagation.

We use the stochastic gradient descent with 0.9 momentum with an annealed learning rate $\mu_p = \frac{0.01}{(1+10p)^{0.75}}$ where p changes from 0 to 1 in the training progress [7, 49] when using LeNet and AlexNet as the classifiers. The learning rate for finetuned layers is set to be the ten percent of that for layers trained from scratch. We use batches with 128 elements in experiments using LeNet, batches with 200 elements in experiments using AlexNet and batches with 36 elements in experiments using ResNet-50.

We use the same architectures and optimization settings (*e.g.*, batch size, learning rate, optimizer and weight decay) as those of the original methods [41, 37] when combining CAT with them.

The pseudo labels are not initialized randomly. Specifically, in the first 5000 iterations, we pre-train CAT by setting $\alpha = 0$. During this, the classifier is trained to fit source data but won't overfit, thus its implicit ensemble can perform well on some target samples and provide a reliable initial set of pseudo labels. Then, we ramp-up to activate the clustering and alignment losses to impose conditional alignment.