# Divide and Adapt: Active Domain Adaptation via Customized Learning Supplementary Materials

Duojun Huang[1,2] Jichang Li[3] Weikai Chen[4] Junshi Huang[5] Zhenhua Chai[5] Guanbin Li[1,2†]
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Research Institute, Sun Yat-sen University, Shenzhen, China
[3]The University of Hong Kong [4]Tencent America [5]Meituan
huangdj9@mail2.sysu.edu.cn, csjcli@connect.hku.hk, chenwk891@gmail.com
{huangjunshi,chaizhenhua}@meituan.com, liguanbin@mail.sysu.edu.cn

We present additional implementation details and analysis of our proposed method DiaNA in this supplementary material.

## 1. Experiment Details

**Implementation details.** We implement all the experiments using PyTorch*. For a fair comparison, most of experimental implementations involving both model training and active learning are consistent with the previous ADA works [3, 6, 7]. The model is initially trained with labeled source data before the sampling steps. During the training phase, we train the model using an Adam [4] optimizer with the learning rate of $10^{-5}$ for DomainNet and an AdaDelta [8] optimizer with the learning rate of 0.1 for Office-Home, while the mini-batch size is set to 64 and 32 for DomainNet and Office-Home, respectively. Furthermore, we set the confidence threshold $\tau$ to 0.95, while we set $k$ used in pairwise feature similarity to 32 and 64 for ResNet-34 and ResNet-50, respectively. Finally, the loss weights $\lambda_c$ and $\lambda_e$ are set to 0.5 and 0.1 respectively.

**Combination with UDA/SSDA/SFDA.** To illustrate the compatibility of the proposed DiaNA with existing UDA/SSDA/SFDA algorithms, we have shown the comparison results in the main text. Here, we give full details on the re-implementations of how to combine DiaNA with these DA methods. Specifically, we initially train the task model with all labeled source samples from $\mathcal{S}$ through supervision. Then, the targeted active samples from the unlabeled target data subset $\mathcal{U}$ would be selected by the proposed sampling strategy, annotated by the human experts, and then moved to the labeled target data subset $\mathcal{T}$. Afterwards, in response to the ADA setup, we conduct domain alignment on $\mathcal{S}$ and $\mathcal{U}$ based on existing DA framework while simultaneously imposing a supervised loss on $\mathcal{T}$ to further optimize the model.

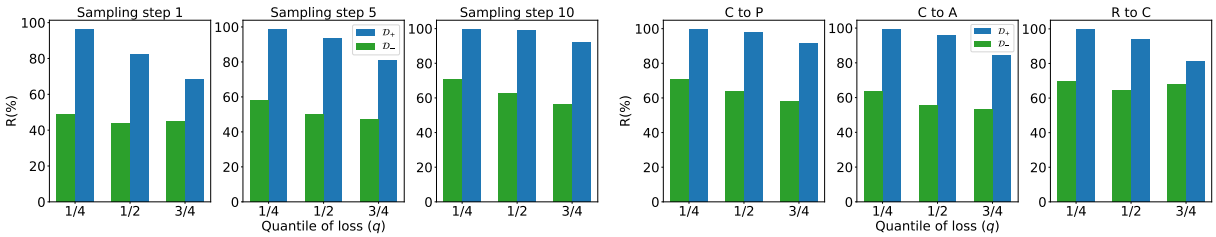
In this work, we need to construct training instances with the aid of labeled samples to obtain the supervision information for training GMM. However, in the SFDA scenario, the labeled source samples cannot be obtained again. Without the aid of labeled examples, this makes it temporarily impossible to divide the target samples into four categories during the first sampling step. Here, we propose to build a model variant of DiaNA to achieve the goal of training the GMM model. Specifically, we first searches for *inconsistent* samples from unlabeled target data and then divide an *uncertain* subset from inconsistent samples by combining the predicted confidence from the model.

Firstly, we select the target samples with high model prediction confidence as the proxy of the labeled data, which can be formulated as follows,

$$\hat{\mathcal{L}} = \{(x, \hat{y}) | \max P(x) \geq t_v, \forall(x,) \in \mathcal{U}\}, \quad (1)$$

where $\hat{y} = \arg\max P(x)$ denotes the predicted class label of the sample $x$ by the model. The confidence threshold value is denoted as $t_v$ to screen confident samples from the target samples. Thanks to the relatively reliable prediction of confident samples, the sample-label pair $(x, \hat{y})$ can be utilized to build the substitutes of the labeled data. As it is necessary to calculate the categorical centroid for each category, the value of $t_v$ is set to 0.95 initially and will be iteratively reduced by 0.1 untill $\hat{\mathcal{L}}$ contains all the categories. After $\hat{\mathcal{L}}$ is obtained, the per-class categorical centroid $A^c$ can be estimated through Eq. (1) with $\mathcal{S}$ replaced by $\hat{\mathcal{L}}$. Further, the similarity-based label $\ddot{y}(x)$ for each unlabeled target sample $x$ is calculated through Eq. (2) in the main text. To obtain the samples with consistent model prediction and similarity-based label and uncertain model prediction ("uncertain-inconsistent") as introduced in the main paper, we obtain the active samples in each sampling step as

†Corresponding Author
*https://pytorch.org/




(a) DomainNet (b) Office-Home

Figure 1. Consistency rates of the well-learnt data subset $\mathcal{D}_+$ and the underfitted data subset $\mathcal{D}_-$ under diverse applications. The hyper-parameter $k$ is set to 32 for ResNet-34 on DomainNet and 64 for ResNet-50 on Office-Home. (a) Across different sampling steps in the adaptation scenario C→S on DomainNet. (b) Across different adaptation scenarios in the first sampling step on Office-Home.

follows,

$$X_3 = \{(x, \hat{y}) | \hat{y} \neq \ddot{y}(x) \textbf{ and } \max P(x) \leq t_c, \forall(x,) \in \mathcal{U}\}, \tag{2}$$

where $t_c$ denotes the confidence threshold value to select uncertain samples. The value of $t_c$ is set to $\frac{1}{C} + 10^{-5}$ initially, where $C$ is the number of categories. Then $t_c$ will be iteratively increased by 0.1 untill $|X_3|$ reaches the annotation quota $b$ in each sampling step. For the following sampling steps, the labeled data will be obtained with $\hat{\mathcal{L}}$ if $A^c$ can not contain all the categories. Otherwise it will be obtained with $\mathcal{T}$ as the formulation Eq. (1).

## 2. Further analysis of DiaNA

### 2.1. Analysis of the top-$k$ similarity

Using categorical centroids and top-$k$ feature similarities, as stated in the main paper, we construct a domainness-based metric to distinguish between source-like samples and target-specific samples from unlabeled target data. In the context of active domain adaptation (ADA), the trained model is inherently biased towards a prominent region of the source domain with high data density [6, 7], as the vast majority of sample labels come from the source. Hence, the source-like target samples, with high feature similarity to the source data, tend to be well learned by the model. On the contrary, as the unique part of target data distribution, the target-specific samples are more likely to suffer from underfitting by the model. Therefore, we aim to construct the domainness metric based on the prediction reliability of sample.

When the value of $k$ in top-$k$ similarity is set to be the full dimension of the feature vector and significantly small, the IoU function in the formulation of similarity based label is equivalent to measuring pairwise image sample similarity under full-resolution and low-resolution conditions, respectively. In the former case, almost all the samples would have identical model predicted class and the label of its closest category prototype in the feature space. When $k$ is set to be significantly small, only the samples with accurate and discriminative features extracted by the model can maintain the consistency between the labels of the two views, since the similarity label is obtained based on only the top-$k$ main components of the sample feature. Here, we make an assumption that when $k$ is set to be small, the well-learned samples tend to maintain a consistent identity for these two labels thanks to their reliable and discriminative features extracted by the model. In contrast, the underfitted samples are more likely to produce a similarity-based label inconsistent with the class label predicted by the model. As the training of model is inevitably dominated by the source domain in the context of ADA, the source-like and target-specific samples in the target well correspond with the well-learned and underfitted samples. Therefore, we utilize the consistency of the model predicted class and similarity-based label to evaluate the domainness of each target sample.

Here, we conduct a validation experiment to investigate the feasibility of our assumption. According to [1, 5], given a trained model, the cross-entropy loss could illustrate how well the model fits the training examples. Hence, we hire it to divide the target samples into the well-learnt and underfitted subsets. Specifically, we should first denote a function to evaluate the loss function value of a target sample $(x,) \in \mathcal{U}$ as follows,

$$\ell(x) = -\sum_{c=1}^{C} \mathbb{1}\{c = y(x)\} \cdot \log P_c(x), \tag{3}$$

where $y(x)$ denotes the actual label of such a sample $x$. After that, we can separate a well-learnt subset and an underfitted subset from all unlabeled target samples as follows,

$$\mathcal{D}_+ = \{x | \ell(x) \leq \ell_q, \forall(x,) \in \mathcal{U}\}, \tag{4}$$

$$\mathcal{D}_- = \{x | \ell(x) > \ell_q, \forall(x,) \in \mathcal{U}\}, \tag{5}$$

where $\ell_q$ is the quantile point of the sorted loss function values. In our implementation, we utilize the 1/4, 1/2, and 3/4 quantile points. Therefore, we define the consistency rate of each data subset as follows,

$$R = \frac{\sum_{x \in \mathcal{D}_*} \mathbb{1}\{\hat{y}(x) = \ddot{y}(x)\}}{|\mathcal{D}_*|}, \tag{6}$$

| Dataset | Office-Home | DomainNet | | |
|---|---|---|---|---|
| Labeling Budget | 5% | 1k | 2k | 5k |
| 1. DiaNA-$\mathcal{L}_{con}$/$\mathcal{L}_{ent}$ for CI | 68.2 / 74.6 | 44.3 / 44.6 | 49.5 / 49.9 | 56.8 / 57.6 |
| 2. DiaNA-$\mathcal{L}_{con}$ for UC/UI/CI | 73.8 / 73.0 / 72.0 | 44.2 / 43.3 / 42.6 | 49.0 / 48.2 / 47.4 | 55.4 / 55.7 / 54.7 |
| 3. DiaNA-$\mathcal{L}_{ent}$ for CC/UI/CI | 74.4 / 74.1 / 73.4 | 44.1 / 44.4 / 44.3 | 47.9 / 48.0 / 47.9 | 54.2 / 53.7 / 53.5 |
| 4. DiaNA | **77.7** | **45.0** | **50.2** | **57.8** |

Table 1. Ablation study of the proposed customized strategy. The performance is evaluated by averaging the accuracy (%) of all the adaptation scenarios.

| Method | A → C | A → P | A → R | C → A | C → P | C → R | P → A | P → C | P → R | R → A | R → C | R → P | AVG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Random | 57.7 | 29.3 | 28.6 | 40.8 | 37.3 | 34.1 | 38.4 | 63.6 | 34.5 | 28.0 | 55.9 | 24.9 | 39.4 |
| BADGE [2] | 63.6 | 49.3 | 46.8 | 64.0 | 50.7 | 50.0 | 71.2 | 72.7 | 51.8 | 68.8 | 72.7 | 53.3 | 59.6 |
| CLUE [6] | 69.5 | 52.9 | 50.0 | 60.8 | 54.7 | 50.5 | 56.0 | 67.3 | 57.7 | 56.8 | 71.8 | 49.3 | 58.1 |
| DiaNA(Ours) | **78.2** | **60.9** | **60.5** | **73.6** | **71.6** | **66.4** | **73.6** | **75.5** | **62.7** | **74.4** | **74.1** | **60.9** | **69.4** |

Table 2. The error rate of the selected samples averaged over all the sampling steps. The experiment is conducted on Office-Home with 5% labeling budget.

where $\mathcal{D}_*$ is a placeholder that represents $\mathcal{D}_+$ or $\mathcal{D}_-$. As shown in Figure 1, when $k$ is set to a value significantly smaller than the feature dimensions, namely 256 for ResNet-34 (or, 2048 for ResNet-50), the consistency rate of the well-learnt data subset is substantially higher than that of the underfitted counterpart. These results demonstrate the feasibility of the proposed assumption to distinguish between the well-learnt source-like samples and the underfitted target-specific samples.

## 2.2. Analysis of customized learning strategy

We conduct more ablation studies to verify the reasonability of our proposed customized learning strategy. For the tailored training objectives designed for different targeted data subsets, we replace the constrained data subset of $\mathcal{L}_{con}$/$\mathcal{L}_{ent}$ with incongruous subsets. As displayed in Table 1 #2-3, DiaNA significantly outperforms all of its variants, demonstrating the superiority of the customized training strategies. In addition, we also withhold CI samples that are incompatible with the current model due to their potential large domain gap with the source domain (see Figure.1(4) in main text). If CI samples are used as constraints, their significant discrepancies between the model's predictions and the similarity labels would jeopardize the training stability. Furthermore, Table 1 #1 reveals the performance of both datasets decreases as a result of additionally adding $\mathcal{L}_{con}$/$\mathcal{L}_{ent}$ to CI samples.

## 2.3. Analysis of the Informative Sampling Function

The data partitioning result produced by the Informative Sampling Function provides a fundamental support for identifying the most informative samples in $\mathcal{U}$. We extensively verify the efficacy of the proposed sampling function constructed based on the Gaussian Mixture Model (GMM). As stated in the main text, the predicted category of each unlabeled target sample in $\mathcal{U}$ is determined by the posterior probabilities of GMM. We further obtain the four-category label according to the piecewise function described in Sec. 3.2 of the main text. The accuracy is calculated as the ratio of the correctly-classified samples in $\mathcal{U}$. It can be observed from Figure 2 that the sampling function is able to identify target samples belonging to the four categories. It should be noted that the accuracy of the model generally increases with the number of labeled images, indicating that the selection and adaptation are complementary to each other for achieving the best domain adaptation performance.

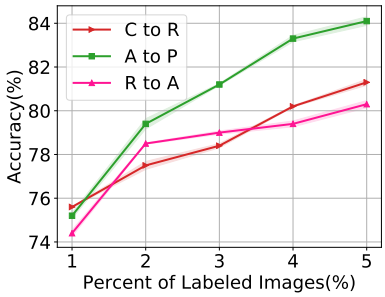



Figure 2. Accuracy of identifying the four categories of unlabeled target data.

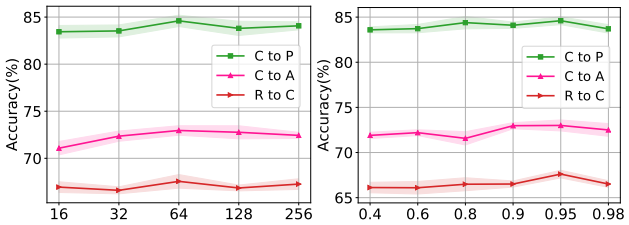
## 2.4. Error rate of the selected samples

The sampling strategy of the proposed DiaNA aims to select the target-specific target samples for annotation, which are typically underfitted by the model as mentioned in Sec. 2.1. To investigate the characteristics of the selected target samples, we report the error rates of all active samples selected by different strategies involving active learning and active domain adaptation. As shown in Table 2, the error rates of DiaNA are consistently higher than the other methods in all cases on Office-Home. This demonstrates that DiaNA is capable of selecting these relatively hard-to-

learn target samples. As illustrated in Figure 2(b) in the main text, annotating these samples and applying supervision to them can potentially correct the model predictions for better adapting to the target data distribution.

### 2.5. Hyper-parameter sensitivity.

We further carry out investigations to check the sensitivity of the proposed approach to the key hyper-parameters $\tau$ and $k$. We conduct the experiments in three diverse adaptation scenarios with varying degrees of transferring difficulty, namely C→P, C→A, and R→C, on Office-Home. In Figure 3, we show the test accuracy of the final adaptation model to display its classification performance when we set $\tau$ and $k$ to $\{0.40, 0.60, 0.80, 0.90, 0.95, 0.98\}$ and $\{16, 32, 64, 128, 256\}$, respectively. As illustrated, the trained model with $\tau = 0.95$ and $k = 64$ together can achieve a relatively higher classification performance than that of other cases, demonstrating the excellent choice of setting both hyper-parameters to 0.95 and 64, respectively.



(a) Analysis of $k$ (b) Analysis of $\tau$

Figure 3. Sensitivity with respect to the hyperparameters of DiaNA.

## References


[1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR, 2019.

[2] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2019.

[3] Bo Fu, Zhangjie Cao, Jianmin Wang, and Mingsheng Long. Transferable query selection for active domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7272–7281, 2021.

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[5] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2019.

[6] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8505–8514, 2021.

[7] Ming Xie, Yuxi Li, Yabiao Wang, Zekun Luo, Zhenye Gan, Zhongyi Sun, Mingmin Chi, Chengjie Wang, and Pei Wang. Learning distinctive margin toward active domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7993–8002, 2022.

[8] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.