Supplementary Material: Understanding and Improving Source-free Domain Adaptation from a Theoretical Perspective

Yu Mitsuzumi[1,2] Akisato Kimura[1] Hisashi Kashima[2]
[1]NTT Corporation [2]Kyoto University
yu.mitsuzumi@ntt.com akisato@ieee.org kashima@i.kyoto-u.ac.jp

A. Formal Definitions and Assumptions

This section is a supplement to the assumptions in Sec. 3.3. To describe the assumptions, we first introduce the formal definition of "expansion".

Definition 1 ((a,d)-expansion)**.** *We say that the class-conditional distribution P_i satisfies (a,d)-expansion if for all $V \subseteq \mathcal{X}$ with $P_i(V) \le a$, the following holds:*

$$P_i(\mathcal{N}(V)) \ge \min\{dP_i(V), 1\} \tag{1}$$

If P_i satisfies (a,d)-expansion for all $\forall i \in [C]$, then we say the distribution P satisfies (a,d)-expansion.

Fig Aa illustrate the visual understanding of (a,d)-expansion.

Using this definition, the formal statements of *expansion* and *separation* assumptions are described as follows.

Assumption 1 (Expansion requirement for unsupervised learning)**.** *We assume that P satisfies $(1/2, d)$-expansion on $\mathcal{X}$ for $d > 1$.*

Assumption 2 (Separation)**.** *We assume P is $\mathcal{B}$-separated with probability $1-\mu$ by the ground-truth prediction model $F^\star$, as follows: $R_{\mathcal{B}}(F^\star) \le \mu$.*

The *expansion* assumption states that the data distribution of the same class is in a continuous region. This is easier to understand if we look at an example that does not satisfy the assumption. As shown in Fig Ab, if the assumption does not hold, there is a small subset V that does not expand with the data augmentation, *i.e.*, $P(V) = P(\mathcal{N}(V))$. This means that the data of the same class are divided into multiple different regions. Conversely, if the assumption holds, the data of the same class are distributed within one region.

The *separation* assumption states that the data of different classes are separated, and the predictions of the ground-truth model among the augmented samples are consistent. This assumption also requires the data augmentation to maintain the semantics of the data.

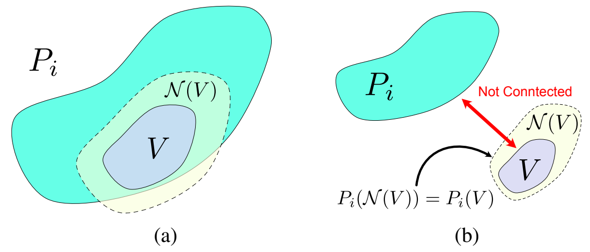


Figure A. **Visual Understanding of Expansion.** (a) Definition of (a,d)-expansion. (b) Example that violates the expansion assumption.

B. Details of Synthetic Data Experiment

In this section, we describe the details of the experiments with the synthetic data in Sec. 3.3. We also describe a similar experiment we conducted using another SFDA method, AaD [1] to confirm the validity of our theoretical analysis.

B.1. Experimental Setups and Detailed Results

Dataset. For the dataset construction, we basically refer to the setting in [1], but we slightly change a few points to see the results more clearly. We construct the source domain data with two inter-twinning moons; in which each moon contains 300 samples. We treat each moon as a discrete class of data. We generate the target domain data by rotating the source domain data by 35° and randomly removing half of the data of one class. The generated dataset is visualized in the leftmost column of Fig. B.

Network Architecture and Training Configuration. The model was a network with three fully connected layers (*2* $\xrightarrow{FC}$ *16* $\xrightarrow{FC}$ *16* $\xrightarrow{FC}$ *2*) as the model. We train it with source domain data by using SGD with a learning rate of $1\mathrm{e}{-2}$, a momentum parameter of 0.9, and a batch size of 64 for 100 epochs. We treated this source-trained model as a source-only one. We subsequently trained the model with the target domain data based on the SHOT-IM loss using SGD with a learning rate of $1\mathrm{e}{-3}$, a momentum parameter

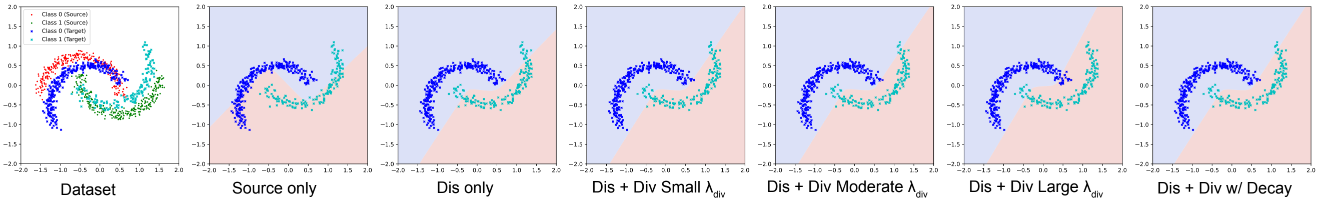


Figure B. **Visualization of the experimental results on SHOT-IM.**

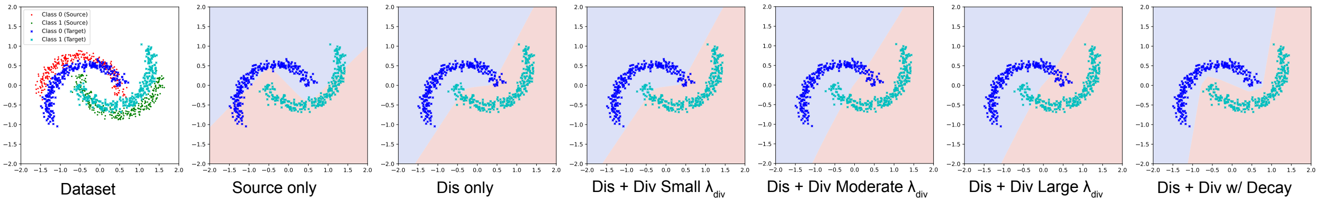
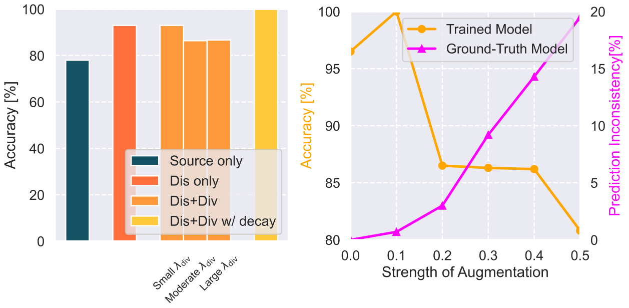


Figure C. **Visualization of the experimental results on AaD.**



(a)

(b)

Figure D. **Experimental Results on Synthetic Data with AaD.** (a) Model accuracy w/ and w/o discriminability and diversity losses. (b) Model accuracy and prediction inconsistency of ground-truth model against the strength of augmentation.

of 0.9, and a batch size of 64 for 500 epochs. We trained the Dis-only model without the diversity loss of SHOT-IM. For the training the Div+Dis model, we set λ_{div} to 0.1, while for training the Div+Dis w/ Decay model, we set λ_{div} to $3.0 \cdot \mathcal{L}_{\text{dis}}$. We set the variance parameter of 2-D Gaussian perturbation used for the data augmentation to 0.05 in the experiment of Tab. 1, while we varied this parameter to obtain Fig. 2.

Detaled results of Tab. 1. For further analysis, we additionally experimented with Dis + Div w/o Decay by varying the weight of the diversity loss. Specifically, we set the weight λ_{div} to 0.05 for "small λ_{div}", 0.1 for "moderate λ_{div}", and 0.3 for "large λ_{div}". The visualized results for Tab. 1 are shown in Fig. B. We can see that a better decision boundary is obtained when the model is trained with both discriminability and diversity losses.

B.2. Additional results using AaD.

Dataset. For the dataset construction, we used the same setting in [1]. We constructed the source domain data with two inter-twinning moons, in which each moon contained 300 samples, and generated the target domain data by rotating the source domain data by $30°$. The generated dataset is illustrated in the leftmost column of Fig. C.

Network Architecture and Training Configuration. We used the same network architecture and source training setup as in the experiment with SHOT-IM. We trained the model with the target domain data based on the AaD loss by using SGD with a learning rate of $1\text{e}{-2}$, momentum parameter of 0.9, and batch size of 64 for 500 epochs. We trained the Dis-only model without the diversity loss of AaD. For training the Div+Dis models, we conducted experiments with several coefficient parameters λ_{div}, namely, we set λ_{div} to 0.1 for small λ_{div}, 0.75 for moderate λ_{div}, and 1.5 for large λ. For training the Div+Dis w/ Decay model, we set the coefficient parameter $\lambda_{\text{div}} = 1.5 \cdot \mathcal{L}_{\text{dis}}$. We set the parameter of the 2-D Gaussian perturbation used for the data augmentation to 0.1 in the experiments, while we varied this parameter in the analysis.

Results. The qualitative and quantitative results are shown in Fig. C and Fig. D. They are mostly consistent with the result of those the SHOT-IM experiment. As shown in Fig. C and Fig. Da, the Dis + Div w/ Decay model performs the best performance, which is in line with the training objective (3). In addition, as shown in Fig. Db, the model accuracy grows as the strength of the augmentation goes large to some extent, but if the strength grows too large, the model accuracy deteriorates as the prediction inconsistency of the ground-truth model increases. The only difference is that in AaD, simply adding the diversity loss does not produce an improvement, which is the reproduced results originally re-

Table A. **Analysis of Augmentation Training (Detailed version of Tab. 5).**

Init. Aug	Aug Training	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.	ΔAaD
Random		**60.1**	78.3	81.8	**70.3**	**80.2**	79.4	66.6	57.9	**82.5**	73.8	60.5	86.1	73.1	+ 0.4
	✓	**60.1**	**78.6**	**82.0**	69.8	**80.2**	**79.9**	**67.1**	**58.3**	82.1	**74.1**	**60.6**	**86.4**	**73.3**	**+ 0.6**
AutoAugment		60.5	**79.3**	**82.1**	69.5	79.3	79.6	**67.6**	58.2	82.2	**74.3**	60.7	86.3	73.3	+ 0.6
	✓	**60.7**	78.9	82.0	**69.9**	**79.5**	**79.7**	67.1	**58.8**	**82.3**	74.2	**61.3**	**86.4**	**73.4**	**+ 0.7**

ported [1]. We suppose that this is because it is inherently important to decay the coefficient of the diversity loss along with the training objective (3). The maximizing marginal entropy loss, which is the diversity loss in SHOT-IM, is a loss in which the gradient magnitude decreases as the training progresses, so it can be regarded as implicitly having a decay mechanism. Therefore, simply adding a diversity loss does not degrade performance in the case of SHOT-IM. On the other hand, the diversity loss of AaD does not work like that of SHOT-IM, meaning that it is necessary to implement a decay mechanism explicitly.

C. Additional Results

Here, we describe the results of the experiments in more detail and provide additional analyses we could not include in the main paper.

C.1. Detailed Results of Analysis

Ablation study of Augmentation Training. Note that we modified a few hyper-parameters ($\eta^{\text{aug}} = 4\text{e}{-3}$, $\lambda_{\mathcal{A}} = 0.25$) when we conducted the analysis with randomly initialized data augmentation. Tab. A shows the detailed results of the ablation study on augmentation training. We can see that the performance with augmentation training is better in 10 out of 12 pairs in the randomly initialized case and 8 out of 12 pairs in the AutoAugment initialized case, demonstrating the validity of our technique.

C.2. Further Analysis of Our Method

Analysis of the coefficient parameter λ_F. As we mentioned in Sec. 4.1, we found in the early study that the heavy augmentations sampled from $\mathcal{A}$ degrade performance, particularly when we use *VisDA2017*. We can mitigate this issue by using the training loss calculated only with weakly augmented samples. Here, λ_F controls the balance between the losses of the heavily augmented samples ($\mathcal{L}_F$) and the weakly augmented samples ($\mathcal{L}'_F$).

In this analysis, we confirmed the validity of using both $\mathcal{L}_F$ and $\mathcal{L}'_F$ by varying the value of λ_F. We used *VisDA2017* for the evaluation, and the other settings were the same as described in Sec 5.1.

The results are shown in Tab. B. When we finely tuned λ_F, the performance of our method reached 88.4%. Using only heavy data augmentations ($\lambda_F = 1.0$) resulted in lower accuracy than when not using them ($\lambda_F = 0.0$). This indicates that using complicated data augmentations does not always improve accuracy. Simply combining $\mathcal{L}_F$ and $\mathcal{L}'_F$ ($\lambda_F = 0.5$) increases accuracy, but by adjusting λ_F appropriately, we can obtain an even greater improvement. Note that, on *Office-31* and *Office-Home*, our method reached sufficient accuracy without any special tuning, so we set λ_F to 0.5 in these cases.

Table B. **Analysis of the coefficient parameter λ_F**

λ_F	Per-class Avg.
0.0 ($\mathcal{L}'_F$ only)	87.6
0.2 (Our setting)	**88.4**
0.5 (Simply combing $\mathcal{L}_F$ and $\mathcal{L}'_F$)	87.8
1.0 ($\mathcal{L}_F$ only)	84.6

References

[1] Shiqi Yang, Yaxing Wang, Kai Wang, Shangling Jui, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. In *Proc. NeurIPS*, 2022. 1, 2, 3