Burn After Reading: Online Adaptation for Cross-domain Streaming Data (Supplementary Material)

Luyu Yang[1], Mingfei Gao[2], Zeyuan Chen[2], Ran Xu[2],
Abhinav Shrivastava[1], Chetan Ramaiah[3]

[1]University of Maryland
[2]Salesforce Research
[3]Google

This supplementary document includes the following:

[A]. An algorithm table of the proposed method.
[B]. Comparison to prior online domain adaptation methods.
[C]. Discussions on streaming randomness.
[D]. Discussions on using other pseudo-label methods.
[E]. A mutli-source setting of the proposed method.
[F]. Hyperparameters.
[G]. Network architecture.
[H]. More ablation study on the usage of augmentations.

A Algorithm Table

We provide an algorithm table of the proposed CroDoBo in Algorithm 1.

B Prior Online UDA approaches

In the main paper, we propose a novel cross-domain framework to implement the right to be forgotten. However, we *do not* claim to have proposed the task of online unsupervised domain adaptation, which has existed before the emergence of deep learning [4,9,17]. The recent works are mostly engineered for a specific downstream task [2,5,26,16,7] that lacks generality. Yet, we try to compare to a more general but unpublished approach CONDA [23] despite its limited availability. The setting of CONDA is different from our approach. It allows a memory module that selectively buffers target queries in which the model can re-access previous target samples. As a result, CONDA is less challenging compared to "burn after reading". Meanwhile, CONDA has a continual setting, in which the model is pretrained on the source domain and then adapted to the target domain. Without any available source code from CONDA [23], we employ the same backbone in [23], *HR-Net* [25], to make a fair comparison. We devise CroDoBoto a continual setting to make it comparable. Without simultaneous access to the source domain, cross-domain bootstrapping is not an option. So we employ the objectives on the target domain (see main paper Sec.3.2.2), we call it

Algorithm 1: The CroDoBo algorithm

Input : Number of learners K, learners $\{\boldsymbol{w}^k\}_{k=1}^K$, learning rate η, number of target queries N_T, confidence threshold τ, batch size B, transform F, data $D_{\mathcal{S}}$, $D_{\mathcal{T}}$, number of class c;

for $j \leftarrow 1$ **to** N_T **do**

  Given $T_j = \{t_b\}_{b=1}^B$ from $D_{\mathcal{T}}$, $\{\tilde{t}_b\} = \{F(t_b)\}$,

  Sample S_j^k from $D_{\mathcal{S}}$, repeat K times;

  for $k \leftarrow 1$ **to** K **do**

    Update $\boldsymbol{w}^k \leftarrow \boldsymbol{w}^k - \eta \nabla \ell_s^k$,

    Obtain pseudo-labels $\{\hat{y}_b^k\}_{b=1}^B = \{\arg\max_c (p(c|t_b; w^k) > \tau)\}_{b=1}^B$;

  end

  for $k \leftarrow 1$ **to** K **do**

    for $b \leftarrow 1$ **to** B **do**

      Obtain $\{\ell_t^{z \to k}\}_{z=1}^{K-1} = \{\mathbf{1}\left(p_b^z \geq \tau\right) \mathcal{H}\left(\hat{y}_b^k; p_b^k\right)\}_{z=1}^{K-1}$,

      Obtain $\ell_{\text{self}} = \ell_{\text{ent}} + \lambda \ell_{\text{div}}$;

    end

  end

  Update $\boldsymbol{w}^k \leftarrow \boldsymbol{w}^k - \eta(\frac{1}{K-1} \sum_{z=1}^{K-1} \nabla \ell_t^{z \to k} + \nabla \ell_{\text{self}})$

  Output $\hat{y}_{\text{test}} = \arg\max_c \frac{1}{K-1} \sum_{k=1}^K p(c|T_j; \boldsymbol{w}^k)$.

end

Continual CroDoBo. The results are in Table 8. We observe that, without any buffer mechanism or re-access to the previous queries, the continual CroDoBo still outperforms ConDA [23].

As mentioned in the main paper Sec.2, we compare to another related task– Test-Time Domain Adaptation [24,22]. We have analyzed the differences of the setting of Test-Time DA in the main paper Sec.2, and here we provide the results of Tent [24] compared with the Continual CroDoBo in Table 9. We observe that our proposed method largely outperforms Tent on VisDA-C.

C Streaming Randomness

As mentioned in the main paper Sec.4, in the online setting, each model takes the same target sequence for fair comparison. The target sequence is randomly-perturbed using the a fixed randomseed. Here, we discuss whether the model will be influenced by different random sequential orders. We perturb the original target sequence (arranged in the categorical order) using 5 different random seeds, and report the results of each seed on VisDA-C [18] and the large-scale Fashion-MNIST-to-DeepFashion [13] benchmark. We compare the randomness using CDAN [15] and CroDoBo. We choose CDAN [15] since it is a benchmark adversarial approach, essentially different from the proposed approach. The results are in Table 10. We observe that on VisDA-C the variance among different sequential orders is rather small (< 0.25). On the more challenging Fashion benchmark, the variance of CroDoBo is larger but manageable ($<$

Table 8. Accuracy on VisDA-C (%) using HR-Net.

Methods (Syn → Real)	plane	bike	bus	car	horse	knife	motor	person	plant	skate	train	truck	Online	One-pass
ConDA [23]	97.0	90.4	80.9	50.0	95.2	95.7	80.3	81.9	94.9	94.2	91.1	63.9	N/P	84.6
Continual CroDoBo (Ours)	96.5	85.2	82.3	47.3	98.0	96.1	89.6	79.2	94.9	95.7	90.4	66.5	80.0	85.1
CroDoBo (Ours)	94.8	86.0	90.7	80.3	97.1	99.1	93.1	85.0	88.2	89.6	90.9	47.1	82.9	86.8

Table 9. Comparisons to *Tent* on VisDA-C using ResNet-101.

Methods (ResNet-101)	plane	bike	bus	car	horse	knife	motor	person	plant	skate	train	truck	Online	One-pass
Tent [24]	85.2	44.3	79.4	50.0	78.1	52.7	83.0	43.5	65.0	53.1	81.4	30.1	62.1	-
Continual CroDoBo (Ours)	93.3	75.8	83.6	70.6	92.8	21.8	86.5	80.5	86.6	90.0	79.6	43.6	74.0	75.4

Table 10. Online accuracy (%) on five different perturbations of target sequence on VisDA-C [18] and Fashion-MNIST [27]-to-DeepFashion [13].

VisDA-C							
Methods	rand 0	rand 1	rand 2	rand 3	rand 4	mean	var
CDAN [15]	62.3	61.0	61.9	61.6	61.9	61.7	0.21
CroDoBo	79.4	78.6	79.6	79.2	79.4	79.2	0.15
Fashion-MNIST-to-DeepFashion							
Methods	rand 0	rand 1	rand 2	rand 3	rand 4	mean	var
CDAN [15]	45.4	47.4	46.7	46.3	46.2	46.4	0.54
CroDoBo	49.1	48.9	46.3	46.5	48.9	47.9	1.99

Table 11. Replacing main paper Eq.3 with other pseudo-labeling methods(%) on VisDA-C.

Methods (Syn → Real)	plane	bike	bus	car	horse	knife	motor	person	plant	skate	train	truck	Online
w^u: MixMatch [1] w^v: FixMatch [20]	93.2	80.9	85.6	67.1	94.1	10.3	88.4	77.9	92.3	91.9	85.7	35.9	74.3
w^u, w^v: MixMatch [1]	94.7	83.3	81.0	62.4	90.7	13.8	84.8	78.7	95.6	94.6	82.9	45.4	71.6
CroDoBo	94.8	87.5	90.5	76.0	94.9	93.7	88.7	80.1	94.8	89.4	84.6	30.7	79.4

Table 12. Performance sensitivity (%) to hyperparameter λ (weight for diversity loss) on VisDA-C [18], τ=0.95.

Metric/λ	0.1	0.4	0.5	0.8	1.0	mean	var
Online	74.9	79.4	78.7	78.5	78.4	78.0	3.1
One-pass	80.2	84.0	83.4	83.6	83.5	82.9	2.4

Table 13. Performance sensitivity (%) to hyperparameter τ (confidence threshold for pseudo-label selection in main paper Eq.2 on VisDA-C [18], λ=0.4.

Metric/τ	0.5	0.6	0.7	0.8	0.9	0.95	mean	var
Online	75.0	76.7	77.3	77.9	78.4	79.4	77.5	2.3
One-pass	80.9	81.7	82.6	82.8	83.4	84.0	82.7	2.0

2.0). We analyze that CroDoBo relies more on the target-oriented supervision (see main paper Sec.3.2.2) than CDAN [15], which makes it more sensitive towards the changes of the target samples. This is a drawback of CroDoBo that we will try to address in the future work.

To conclude, the randomness in forming the order of target queries will not be a factor that influences the evaluation of the online model effectiveness.

D Other Pseudo-labeling Approaches as Co-supervision

The co-supervision in the proposed method (cf. main paper Eq.3) can be replaced with any other pseudo-labeling approaches. One can simply replace the term on either/both $\{w^u, w^v\}$ to achieve better performance. We replace on either/both learners with another popular semi-supervised approach MixMatch [1] and report the results in Table 11. We observe that FixMatch [20] provides better co-supervision and the online performance drops $\sim$8% when replaced with MixMatch.

E Multi-Source CroDoBo.

Since the proposed method exploits the learners' discrepancy, a natural extension of the proposed method is to use multiple source to obtain more discrepant co-supervisions. We experimented on VisDA-C with one learner taking from an additional source domain from the Youtube Bounding Box dataset [18]. For fair comparison, we randomly select a subset of the source samples to have equal total number of source samples for both multi-source and single-source settings. The results are reported in Table 14. Multi-Source CroDoBo improves the class average accuracy to a remarkable 87.8%. The result further validates the effectiveness to increase data diversity.

F Hyperparameters

We have two hyperparameters in the proposed approach: λ for weighing the term ℓ_{div} and τ for the pseudo-label selection (cf. main paper Eq.2). We used λ=0.4 and τ=0.95 in all our experiments, here we report results on more settings of these hyperparameters. The results of λ=$\{0.1, 0.4, 0.5, 0.8, 1.0\}$ are shown in Table 12. As the results suggest, CroDoBo is not sensitive to hyperparameter λ. We observe similar performance of the model when λ is larger than 0.4.

Table 14. Accuracy on VisDA-C (%) with Multi-Source CroDoBo$^+$.

Methods (ResNet-101)	plane	bike	bus	car	horse	knife	motor	person	plant	skate	train	truck	Online	Class Avg.
Multi-Source CroDoBo$^+$	96.2	85.4	90.8	79.7	96.6	94.6	93.3	87.5	96.3	92.4	90.2	50.6	84.0	87.8
CroDoBo$^+$ K=2	94.8	87.5	90.5	76.0	94.9	93.7	88.7	80.1	94.8	89.4	84.6	30.7	79.4	84.0

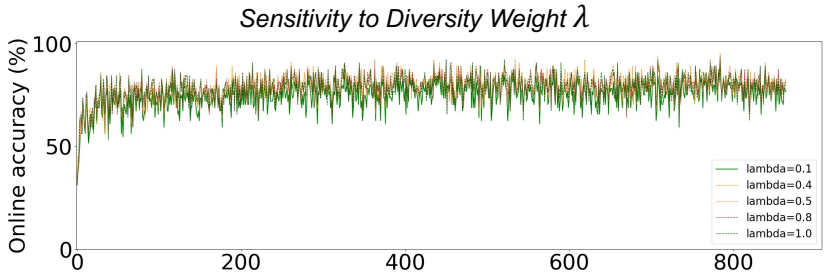

Fig. 7. Results of online accuracy *w.r.t.* sensitivity to hyperparameter λ for the diversity term on VisDA-C [18] using ResNet-101.

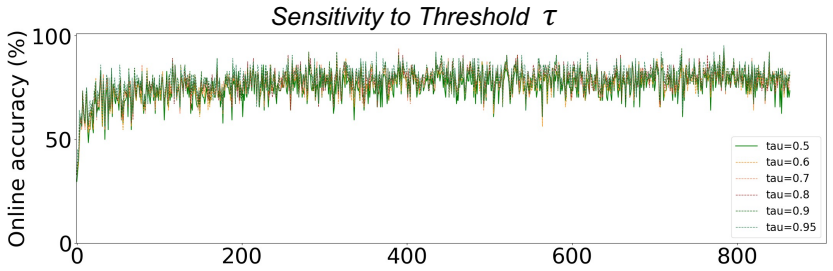

Fig. 8. Results of online accuracy *w.r.t.* sensitivity to hyperparameter τ for pseudo-label threshold on VisDA-C [18] using ResNet-101.

The sensitivity to τ is shown in Table 13. When τ is smaller, more samples in each target query are selected as pseudo-labels to co-supervise the peer learner. However, the quality of these pseudo-labels is compromised since the model is less confident about the prediction. Thus, the co-supervision is less accurate to depend on. We observe the performance drop when the threshold τ is smaller than 0.6. Therefore, we suggest a larger threhold τ to achieve a more effective model. The online accuracy of the above settings are shown in Figure 7 and Figure 8.

G Network Architecture

We follow the network architecture in [11,12], a feature backbone followed by a bottleneck layer with dimension=256, and a Linear layer as the output layer. For the experiments on VisDA-C [18], COVID-DA [28] and Fashion-MNIST-to-DeepFashion [27,13], the feature backbone is pretrained on ImageNet [3]. For the *WILDS*-Camelyon17 benchmark, we followed the leaderboard to use a randomly initialized DenseNet-121 [8]. We use Adam [10] with with an initial learning rate of 8e-4. The query size in our experiments is set as 64. We have not observed

Table 15. Ablation study of *RandAug* and the multiple forward of each target query on VisDA-C and Fashion dataset.

Dataset	VisDA-C		Fashion	
	No Aug	*RandAug*+multiple use of target	No Aug	*RandAug*+multiple use of target
DAN [14]	57.8	68.4	40.7	45.2
CORAL [21]	66.7	72.1	40.4	37.1
DANN [6]	49.0	49.9	35.6	37.2
ENT [19]	55.8	46.1	31.9	31.3
MDD [29]	60.4	67.0	36.5	39.0
CDAN [15]	62.3	62.8	45.4	41.0
CroDoBo K=2 (Ours)	**77.9**	**79.4**	**47.6**	**49.1**

any major performance change using different batch-size. Results are reported based on an average of 5 runs.

H More Ablation Study

As clarified in Sec.3.2, *RandAug* is only employed to increase the data diversity, and is not required for the proposed method. We note that the use of *RandAug* and the multiple use of each target query in the proposed method might lead to confusion. To better evaluate the proposed method, besides providing CroDoBo without any augmentation in the main paper, here we further provide the augmented baseline results, and with multiple use of each target query with two strong and two weak augmented versions. We search the best performing hyperparameters for each method using grid-search. We observe that (Table 15) either CroDoBo or CroDoBo+ outperforms the compared baselines.

References

1. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249 (2019) 3, 4
2. Delussu, R., Putzu, L., Fumera, G., Roli, F.: Online domain adaptation for person re-identification with a human in the loop. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 3829–3836. IEEE (2021) 1
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 5
4. Dredze, M., Crammer, K.: Online methods for multi-domain learning and adaptation. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. pp. 689–697 (2008) 1
5. Gaidon, A., Vig, E.: Online domain adaptation for multi-object tracking. arXiv preprint arXiv:1508.00776 (2015) 1
6. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR **17**(1), 2096–2030 (2016) 6

7. Hajifar, S., Sun, H.: Online domain adaptation for continuous cross-subject liver viability evaluation based on irregular thermal data. IISE Transactions pp. 1–12 (2021) 1
8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017) 5
9. Jain, V., Learned-Miller, E.: Online domain adaptation of a pre-trained cascade of classifiers. In: CVPR 2011. pp. 577–584. IEEE (2011) 1
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 5
11. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International Conference on Machine Learning. pp. 6028–6039. PMLR (2020) 5
12. Liang, J., Hu, D., Feng, J.: Domain adaptation with auxiliary target domain-oriented classifier. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16632–16642 (2021) 5
13. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1096–1104 (2016) 2, 3, 5
14. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International conference on machine learning. pp. 97–105. PMLR (2015) 6
15. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. arXiv preprint arXiv:1705.10667 (2017) 2, 3, 4, 6
16. Mancini, M., Karaoguz, H., Ricci, E., Jensfelt, P., Caputo, B.: Kitting in the wild through online domain adaptation. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1103–1109. IEEE (2018) 1
17. Moon, J., Das, D., Lee, C.G.: Multi-step online unsupervised domain adaptation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 41172–41576. IEEE (2020) 1
18. Peng, X., Usman, B., Kaushik, N., Wang, D., Hoffman, J., Saenko, K.: Visda: A synthetic-to-real benchmark for visual domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 2021–2026 (2018) 2, 3, 4, 5
19. Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K.: Semi-supervised domain adaptation via minimax entropy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8050–8058 (2019) 6
20. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685 (2020) 3, 4
21. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. In: AAAI (2016) 6
22. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International Conference on Machine Learning. pp. 9229–9248. PMLR (2020) 2
23. Taufique, A.M.N., Jahan, C.S., Savakis, A.: Conda: Continual unsupervised domain adaptation. arXiv preprint arXiv:2103.11056 (2021) 1, 2, 3
24. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2020) 2, 3

25. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence (2020) 1
26. Wu, D.: Online and offline domain adaptation for reducing bci calibration effort. IEEE Transactions on human-machine Systems **47**(4), 550–563 (2016) 1
27. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017) 3, 5
28. Zhang, Y., Niu, S., Qiu, Z., Wei, Y., Zhao, P., Yao, J., Huang, J., Wu, Q., Tan, M.: Covid-da: Deep domain adaptation from typical pneumonia to covid-19. arXiv preprint arXiv:2005.01577 (2020) 5
29. Zhang, Y., Liu, T., Long, M., Jordan, M.: Bridging theory and algorithm for domain adaptation. In: International Conference on Machine Learning. pp. 7404–7413. PMLR (2019) 6