# Hard Class Rectification for Domain Adaptation

**Ref. No.: KNOSYS-D-20-02109**

Dear Editor and Reviewers,

Please find enclosed our revision to Ref. No.: KNOSYS-D-20-02109, “Hard Class Rectification for Domain Adaptation”. We have revised the paper, considering the reviewer’s comment.

We thank the reviewers for their very helpful and detailed comments and hope that we addressed all of their concerns in this revision. It is noteworthy that the modified text is marked in red in the revised manuscript.

Sincerely,

Yunlong Zhang, Changxing Jing, Huangxing Lin, Chaoqi Chen, Yue Huang, Xinghao Ding, Yang Zou

## 1 Response to Reviewer 1

**Comment:** This paper proposes a hard class rectification pseudo-labelling for domain adaptation. Specifically, an adaptive prediction calibration (APC) scheme is proposed to identify and calibrate the target hard classes. For the hard classes, temporal resembling and self-ensembling are used to achieve the prediction consistency. Extensive experiments are conducted to evaluate the proposed method. In general, the novelty of the method is limited. The idea of exploring hard and easy classes has been proposed in previous domain adaptation methods (e.g. CBST, etc). However, it fails to discuss the key differences between the proposed method and the previous ones. Besides, the technical contribution is also limited, since most of the components are borrowed from the standard semi-supervised learning methods (e.g. ReMixMatch, $\pi$-model, and Mean Teacher) and domain adaptation methods (e.g. CBST). It is okay that the proposed method is a combination of previous components. However, it would be better if more insightful justification is discussed on why these components are complementary to each other and can jointly contribute to the current task.

Thank you for your suggestion. We first illuminate the novelty of this paper more clearly. *It is noteworthy that the "hard class" mentioned in CBST is close to the normal class mentioned in this paper.* Here, we further introduce the easy, normal, and hard classes proposed in this paper. As shown in Figure 1, the classifier is trained on the source domain, and three classes deliver distinct performances. Class 1 belongs to easy classes and has a higher predictive class proportion (i.e., (the number of samples classified into a certain class)/(the number of target samples)). The target samples belonging to these classes are very likely to be classified correctly. The classifier has good generalization ability to easy classes. Class 2 belongs to normal classes and has a moderate predictive class proportion. Although a part of the target samples belonging to it is misclassified, the predictions with higher confidence have higher accuracy. Therefore, the existing pseudo-labeling methods progressively improve the performance of these classes by adding the samples with higher confident predictions into training. Class 3 belongs to hard classes and has a lower predictive class proportion. Meanwhile, for hard classes, even the target samples with higher confident predictions also are highly possible to be misclassified, adding these samples into training will misguide the classifier. Therefore, the pseudo-labeling methods cannot improve the performance of the hard classes even will deteriorate it, which is the main difference between the normal and hard classes. As far as we knew, this paper is the first work to consider this problem. The novelty in this paper is revealing the hard class problem neglected in the former work. This analysis is shown in the fourth paragraph of Sec 1 in the revised manuscript.

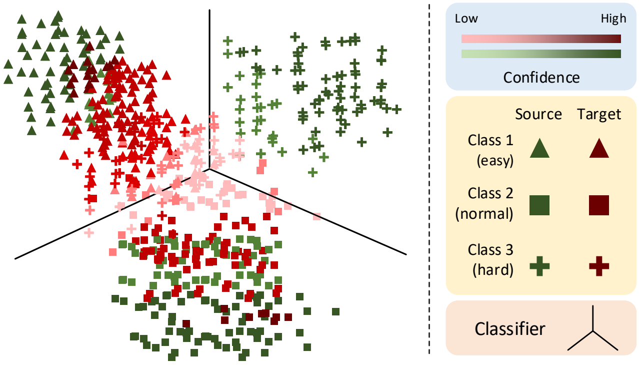
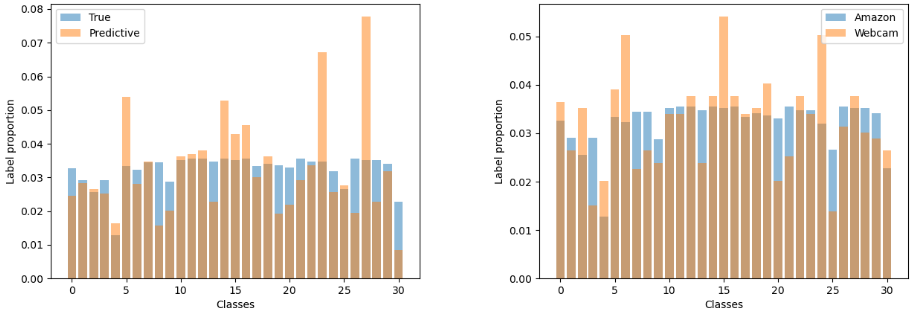


Figure 1: Hard class problem in existing pseudo-labeling based DA methods: Compared with class 1 and 2, class 3 has lower predictive class proportion. Meanwhile, for this class, target samples with higher confidence are mainly classified into class 1.



(a)

(b)

Figure 2: (a) The difference of predictive and true label proportions; (b) The difference of label proportion for source and target domains.

**Comment:** In the following [1], a very similar calibration method using the prior distributions is proposed. So what is the difference between the proposed one and the one used in [1]? And why it also works here?

Thank you for pointing this out. We first answer the question that why the APC works. We propose the Adaptive Prediction Calibration (APC) due to the hard class problem, which leads to a mismatch between predictive and true label proportions, which is shown in Figure 2(a). The APC is proposed to eliminate this mismatch between the predictive and true class proportion. However, the true class proportion of the target domain is unknown. Here, the label proportion of the source domain is used to estimate the prior class proportion, which is reasonable due to the following two aspects. First, the label proportions of the source and target domains always are close (e.g. Figure 2(b)), which is valid in many practical applications. Second, although a minor disagreement between the label proportions of source and target domain, the resulting performance is slightly affected. For example, we respectively utilize the marginal class distribution of the source and target domains as prior class proportion and report the results in Table 1. The results illuminate that the transfer performances are close between 'S' and 'T'.

We further explain the difference between the proposed one and the one used in [1]. Most importantly, we introduce the APC due to the mismatch between predictive and true label proportions. If without extra constraint, pseudo-labeling may be misguided by massive false pseudo labels for hard classes. In comparison, this problem is not evident in semi-supervised learning. Therefore, it is

Table 1: Comparison of different prior class proportion on Office-31 under UDA setting (%). S denotes the marginal class proportion of the source domain. T denotes the marginal class proportion of the target domain.

| prior class proportion | A → W | D → W | W → D | A → D | D → A | W → A | Avg |
|---|---|---|---|---|---|---|---|
| S | 95.9±0.2 | 98.7±0.1 | **100.0**±0.0 | 94.3±0.2 | 75.0±0.4 | 75.4±0.4 | 89.9 |
| T | **96.3**±0.1 | **98.9**±0.1 | **100.0**±0.0 | **94.9**±0.2 | **75.5**±0.3 | **76.2**±0.3 | **90.3** |

more reasonable to use it here. Otherwise, we adopt the label proportion of source domain as prior class proportion, which is a more accurate evaluation than the calibration method in [1].

**Comment:** The idea of TE and SE in this paper are based on $\pi$-model and Mean Teacher. However, the difference between the proposed method and $\pi$-model and Mean Teacher is not well justified. Besides, What is the motivation for using them? What is the basis? "I did it and it worked"-kind of reasoning is not a suitable mechanism to advance science.

Thank you for pointing this out. The consistency constraints have been fully used in semi-supervised learning and domain adaptation. Here, we have strong reasons to use them, that is *the APC will further magnify the unrobustness of predictions for hard class*. Concretely, the predictions of out-of-distribution samples are vulnerable [2]. Furthermore, the predictions of samples belonging to the hard class are highly-magnified, which results in the fact that the vulnerability of predictions is also magnified. Therefore, it is more urgent to improve the robustness of predictions for the model with the APC. In this paper, we proposed the TE and SE to tackle this problem. This analysis also is shown in the fifth paragraph of Sec 1 in the revised manuscript.

**Comment:** The writing needs some improvements. There are many typos and grammatical issues. To name a few, in the abstract: label-scare → label-scarce, hard class → hard classes

Thank you for pointing this out. We have carefully checked and corrected the spelling and grammar of the full paper, especially for the mistakes pointed out by the reviewer.

## 2 Response to Reviewer 2

**Comment:** According to this paper, it reveals the hard class problem in the domain adaption area. However, the hard class problem is not well defined. More specifically, the differences between easy class problem, normal class problem, and hard class problem are not quantitatively defined in this paper.

Thanks for the comments. Here, we define the easy, normal, and hard classes more clearly. As shown in Figure 1, the classifier is trained on the source domain, and three classes deliver distinct performances. Class 1 belongs to easy classes and has a higher predictive class proportion (i.e., (the number of samples classified into a certain class)/(the number of target samples)). The target samples belonging to these classes are very likely to be classified correctly. The classifier has good generalization ability to easy classes. Class 2 belongs to normal classes and has a moderate predictive class proportion. Although a part of the target samples belonging to it is misclassified, the predictions with higher confidence have higher accuracy. Therefore, the existing pseudo-labeling methods progressively improve the performance of these classes by adding the samples with higher confident predictions into training. Class 3 belongs to hard classes and has a lower predictive class proportion. Meanwhile, for hard classes, even the target samples with higher confident predictions also are highly possible to be misclassified, adding these samples into training will misguide the classifier. Therefore, the pseudo-labeling methods cannot improve the performance of the hard classes even will deteriorate it, which is the main difference between the normal and hard classes. As far as we knew, this paper is the first work to consider this problem. The novelty in this paper is revealing the hard class problem neglected in the former work. This analysis is shown in the fourth paragraph of Sec 1 in the revised manuscript.

**Comment:** There are some problems with the structure of the paper. The 4th part theoretical analysis seems to have no straight link to both methods and experiments part. It is recommended to move the theoretical analysis part before the methods part or to add it to the beginning of the methods part.

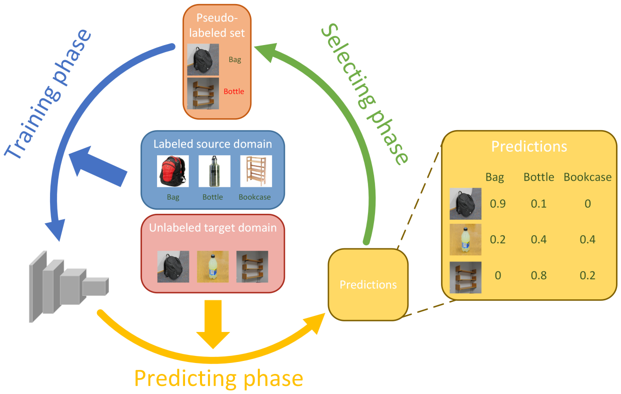


Figure 3: The overall algorithm structure block diagram

Thank you for pointing this out. We have adjusted the structure of the article. Concretely, we have increased a new paragraph (Sec 3) that describes the hard class problem and added Sec 4 into it to emphasize that the hard classes harm the DA in the revised manuscript.

**Comment:** There isn't an overall algorithm structure block diagram, which is not convenient and intuitive if the readers want to grasp the idea of the whole paper. Therefore, it is recommend to draw an overall algorithm structure block diagram.

We have added the overall algorithm structure block diagram (Fig 2 in the revised manuscript.) to improve the readability of papers. It also is shown in Figure 3.

**Comment:** It is not mentioned in this paper whether HCRPL has any restricts or extra requirements, which is quite concerning. The strengths and limits of the new method are not specifically discussed in this paper, which makes the paper incomplete.

Thank you for pointing this out. We have further discussed the strength and weaknesses of our approach in the revised manuscript (Sec 6). We also copy it here.

**Strength** The major contribution of this paper is unraveling the hard class problem, which always is ignored in the existing pseudo-labeling methods yet is critical in some practical scenarios. Compared with the existing pseudo-labeling methods, our approach improves not only the overall performance but only the worst performance among all classes.

**Weakness** The introduction of prior knowledge alleviates the hard class problem effectively. However, it also limits the application of our approach to a certain extent. Our basic assumption is that the source and target domains should have similar label proportion, which is invalid in some applications, such as partial domain adaptation [3], open set domain adaptation [4]. In such variants of DA, applying APC obviously will miscalibrate the predictions and results in massive false pseudo labels. Therefore, accurately inferring the marginal class distribution of the target domain should be further studied in the future job. On the other hand, the proposed HCRPL can effectively improve the precision of hard classes but has little impact on the recall of them, which also is important in practical applications.

## 3 Response to Reviewer 3

**Comment:** There are some grammar errors or typos in the manuscript. For example, in abstract, "As [it] is . . . ", "as mention[ed] before". Please check carefully.

We have carefully checked and corrected the spelling and grammar of the full paper, especially for the mistakes pointed out by the reviewer.

**Comment:** What is the meaning of global alignment? Besides, is there any evidence or literature that demonstrates the proposed two limitations? Actually, there are many works that deal with conditional distribution alignment using the pseudo label idea, such as [A,B,C]. Please clarify their differences.

The global alignment means the alignment of the source and target domains in the latent feature space. Recently, some literature [5–10] demonstrates either or both of the proposed limitations harms the performance of the methods based on global alignment.

The negative effects resulting from the false labels are critical in the pseudo-labeling. Hence, we further clarify the difference between [A,B,C] from the aspects of improving the accuracy of pseudo labels. [A] combined the pseudo-labeling with re-weighting. The latter can alleviate the results of the negative effects from the false pseudo labels. [B] improved the accuracy of pseudo labels by integrating the predictions of various classifiers. [C] improved the quality of the pseudo labels by refining the labels after each iteration. This analysis is shown in Sec 2 in the revised manuscript.

**Comment:** What is the meaning of false pseudo-labels ratio? Also, the content in Section 4 is hardly called theoretical analysis in this paper. The only theory is from existing work [2].

The false pseudo-label ratio can be calculated by (the number of incorrectly pseudo-labeled samples)/(the number of pseudo-labeled samples). Otherwise, we have deleted the theoretical analysis and moved Section 4 after Sec 2 for rigor in the revised manuscript.

**Comment:** Please carefully use the word "To the best of our knowledge, . . . first time."

We have deleted such expression and replaced it with "in this paper, we reveal the hard class problem in domain adaptation.", which appears in the summary in Section 1.

**Comment:** Domain adaptation is a hot issue in recent years and lots of approaches are proposed. Please list the main differences between the proposed method and the existing methods.

As the reviewer said, DA is a hot issue in recent years. As the recent survey [11], the DA methods can be grouped into several categories: Domain-Invariant Feature Learning, Domain Mapping, Normalization Statistics, Ensemble Methods, and Target Discriminative Methods. The proposed HCRPL is based on pseudo-labeling, which is included in ensemble methods. Compared with the other related methods, we found that these methods may deteriorate the worse performance among all classes, which is called the hard class problem. To alleviate this problem, we propose the APC, SE, and TE.

**Comment:** The difference of TE and SE between this paper and [14, 6, 37] is unclear. The authors mentioned that the aim is different, how about the actual model? It seems that they are the same, just used in a different situation, like existing model A works for images and you use it for text?

Thank you for pointing this out. The consistency constraints have been fully used in semi-supervised learning and domain adaptation. Here, we have strong reasons to use them, that is *the APC will further magnify the unrobustness of predictions for hard class*. Concretely, the predictions of out-of-distribution samples are vulnerable [2]. Furthermore, the predictions of samples belonging to the hard class are magnified proportionally, which results in the fact that the vulnerability of predictions is also magnified. Therefore, it is more urgent to improve the robustness of predictions for the model with the APC. In this paper, we proposed the TE and SE to tackle this problem. This analysis is shown in fifth paragraph of Sec 1 in the revised manuscript.

**Comment:** Important references [A-F] are missing, please make comments and compare with them if possible.

We have cited and discussed these papers in the revised manuscript as follows.

To aligning conditional distribution, many pseudo-labeling based DA methods [12–14] are proposed. [A] [12] introduced the label propagation to update the pseudo-labels and proposed landmark selection to re-weight the samples of source and target domains. [B] [13] determined the pseudo labels by integrating the predictions of multiple classifiers. [C] [14] improved the quality of the pseudo labels by refining the labels after each iteration.

Furthermore, we compare the proposed method with [A], [C], [D], [E], and [F] in the experiments. This analysis and results are shown in Sec 2 and Table 1-4 in the revised manuscript.

**Comment:** How to choose the parameters? Indeed, the authors should perform cross-validation to pick up the parameters.

Under our experimental setup, it is impossible to tune the optimal parameters using cross-validation, since labeled and unlabeled data are sampled from different distributions. Thus, we evaluate our approach by empirically searching the parameter space for the optimal parameter settings on Office-31 $a \rightarrow w$ under UDA setting and then apply the optimal parameters in all experiments. Moreover, the main hyper-parameters in this paper are EMA momentum $\alpha$ and sharpening temperature $T$. We have designed the sensitivity analysis for them on $A \rightarrow W$. The results is shown in Table 2. According to the results that the HCRPL achieves the best performance when $\alpha = 0.95$ and $T = 0.5$, we confirm the hyper-parameters and apply them on all settings. The table and analysis are shown in Table 8 and Sec 5.6.6 in the revised manuscript.

Table 2: Comparison of different EMA momentum $\alpha$ and sharpening temperature $T$ on A→W.

| $\alpha$ | 0.0 | 0.8 | 0.9 | 0.95 | 0.97 | 0.99 |
|---|---|---|---|---|---|---|
| Acc (%) | 92.5 | 93.1 | 95.2 | **95.9** | 94.5 | 93.1 |
| $T$ | 0.2 | 0.3 | 0.5 | 0.7 | 0.8 | 1.0 |
| Acc (%) | 94.7 | 95.7 | **95.9** | 93.8 | 93.8 | 92.3 |

**Comment:** The Office-31 data set is quite a toy data here. Actually, existing methods have shown that the learning tasks in this data set are relatively easy. Please add the Office-Caltech data set, which is more difficult compared with the Office-31. Besides, [A,C,E] may also have experiments on the Office-Caltech data set, please compare with these methods.

We have added the results and related analysis of Office+Caltech in the Subsection 5.4.1 of revised manuscript. Here, we also report the results in Table 3. In total, our approach outperforms the [A,C,E] and some other existing methods.

**Comment:** As the proposed method is based on deep learning, the performance may be an important estimation of the model. However, the compared baseline methods seem out-of-date. Please compare with newly released domain adaptation methods, especially those published in 2020. Besides, as mentioned before, there are lots of domain adaptation methods, make sure to compare with a few classical approaches, such as TCA [F].

In experiments, we further add some classical and latest methods as compared baseline methods, such as TCA [17], SA [15], GFK [16], ATM [23], GSDA [24], AADA [25]. We compare the proposed HCRPL with them on Office-31, Office+Caltech, ImageCLEF-DA, and Office-Home, which is shown in the revised manuscript (Table 1-4). The experiments demonstrate that our approach is better than the traditional methods by a large margin and also outperforms the latest methods.

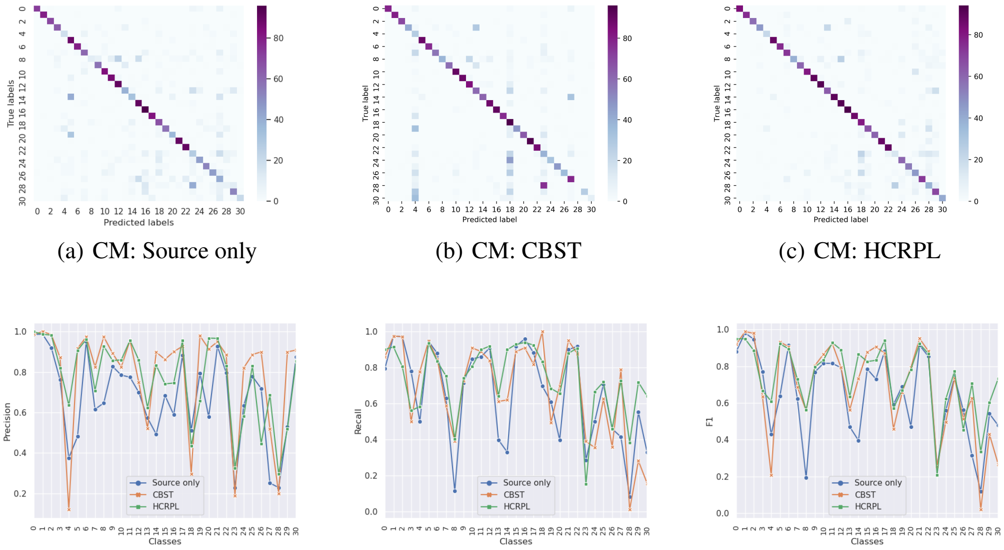
**Comment:** Figure 5 uses the Office-31 data set as an example, as mentioned before, the learning tasks here may not be difficult. I suggest the authors use another data set in experiments. The same issue occurs in Figure 6.

Office-31 Webcam → Amazon is a quite challenging task, and the test accuracy achieves 60.7 and 75.4 using source only and HCRPL. Meanwhile, this task can confirm the existence of the hard class problem. Hence, we set it as an example to study the hard class problem in-depth.

**Comment:** In Section 5.6.2, "We further analyze the result shown in Figure 5 and consider the class with worst performance, 'Stapler'. As shown in Figure 7," what do you mean? Figure 5 or Figure

Table 3: Shallow and deep approaches on Office+Caltech under UDA setting (%)

| S → T | SA [15] | GFK [16] | TCA [17] | SCA [18] | LPJT [19] | KJDIP-rbf [14] | AlexNet [20] | MMD-CORAL [21] | GKE [22] | HCRPL |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Shallow | | | | Deep | | |
| A→ C | 80.1 | 76.9 | 74.2 | 78.8 | 85.4 | 85.8 | 84.6 | **89.1** | 88.4 | **89.1** |
| A→ D | 78.3 | 79.6 | 78.3 | 85.4 | - | 87.9 | 88.5 | 96.6 | **99.7** | 95.8 |
| A→ W | 68.8 | 68.5 | 71.9 | 75.9 | 92.2 | 91.2 | 83.1 | 95.7 | **97.6** | 95.9 |
| C→ A | 89.5 | 88.4 | 89.3 | 89.5 | 92.1 | 92.4 | 91.8 | 93.6 | 93.5 | **94.0** |
| C→ D | 83.4 | 84.6 | 83.4 | 87.9 | - | 90.4 | 89.0 | 93.4 | 94.3 | **98.3** |
| C→ W | 75.9 | 80.7 | 80.0 | 85.4 | 92.7 | 89.5 | 83.1 | 95.2 | 98.3 | **98.8** |
| D→ A | 82.7 | 85.8 | 88.2 | 90.0 | - | 89.4 | 89.3 | 94.7 | 93.5 | **94.8** |
| D→ C | 75.7 | 74.1 | 73.5 | 78.1 | - | 78.5 | 80.9 | 84.7 | 83.8 | **89.2** |
| D→ W | 99.3 | 98.6 | 97.3 | 98.6 | - | 97.6 | 97.7 | 99.4 | **99.7** | 99.6 |
| W→ A | 77.8 | 75.3 | 80.0 | 86.1 | 92.3 | 92.1 | 83.8 | **94.8** | 94.4 | 94.6 |
| W→ C | 74.9 | 74.8 | 72.6 | 74.8 | 86.0 | 83.5 | 77.7 | 86.5 | 88.9 | **88.9** |
| W→ D | **100.0** | **100.0** | **100.0** | **100.0** | - | 96.8 | **100.0** | **100.0** | **100.0** | **100.0** |
| Avg | 82.2 | 82.4 | 82.4 | 85.9 | - | 89.6 | 87.5 | 93.6 | 94.3 | **94.9** |



(a) CM: Source only (b) CM: CBST (c) CM: HCRPL

(d) Precision (e) Recall (f) F1-score

Figure 4: (a)-(c) The Confusion Matrix (CM) visualization for Source only, CBST, and HCRPL. (d)-(f): The precision, recall, and f1-score evaluated on three different models, Source-only, CBST, and HCRPL. The result is obtained on Office-31 W → A under UDA setting using ResNet-50. To better visualize the results, we arrange the categories in alphabetic order.

7, it looks like there are typos here. Besides, where is the Stapler come from? How to measure the performance? The statements are far from clear.

Thank you for pointing this out. To avoid the ambiguity resulting from Figures 5 and 7, we have modified and merged these two figures and reported the confusion matrix, precision, recall, and f1-score of the source only, CBST, and our approach in Figure 6 in the revised manuscript. Here, we also report the results in Figure 4. Among them, the confusion matrix shows the detailed results, and the others are three common class-level metrics. We first compare the results of Source only and CBST. Although the precision, recall, and f1-score of majority classes are better to a certain extent, the worse performances among all classes may deteriorate further. Here, we focus on the performance of 28-th class in the confusion matrix. The majority of samples belonging to this class are misclassified into the 23-th class in the case of training on the source domain only. After pseudo-labeling, the predictions more center on the 23-th class, which results from the misguidance of false predictions with high confidence. The results of precision, recall, and f1-score also confirm this conclusion. Furthermore, we found that the predictive class proportion of easy classes, such as 4-th and 18-th classes, are higher after applying pseudo-labeling, which results in the decline of precision for these classes, which also is the drawback of pseudo-labeling.

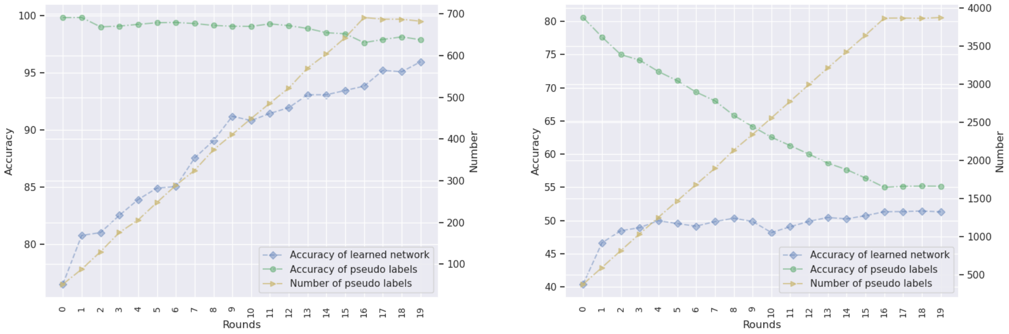
We further analyze the performance of our HCRPL from the aspect of hard classes. First, the HCRPL avoids the predictions degenerate into few classes, which illustrates the HCRPL can control the predictive class proportion of each class in a reasonable interval and improve the precision of the easy classes. Second, we found the precision of majority classes, is better than CBST and source only. The recall is flat or slightly better than CBST and source only. Overall, class-level performance exceeds CBST and source only. Especially for hard classes, such as 28-th, and 30-th classes, the performance of HCRPL is superior apparently, which demonstrates that our approach can indeed alleviate the hard class problem.

**Comment:** These related works should be included in the references:

[A]. Locality preserving joint transfer for domain adaptation [J]. IEEE TIP 2019.

[B]. Iterative Refinement for Multi-source Visual Domain Adaptation [J]. IEEE TKDE, 2020.

[C]. Domain adaptation by joint distribution invariant projections [J]. IEEE TIP 2020.



(a) Office-31 Amazon→Webcam

(b) Office-Home Product→Clipart

Figure 5: (a)-(b): Comparison of the actual accuracy of pseudo-labels and learned network accuracy during training.

[D]. Informative Feature Selection for Domain Adaptation [J]. IEEE Access 2019.

[E]. Geometric Knowledge Embedding for unsupervised domain adaptation [J]. KBS 2020.

[F]. Domain adaptation via transfer component analysis [J]. IEEE TNN 2010.

[G]. Collaborative Unsupervised Domain Adaptation for Medical Image Diagnosis, IEEE TIP, 2020

[h]. Semi-Supervised Optimal Transport for Heterogeneous Domain Adaptation, IJCAI, 2018

Thank you for pointing this out. We have cited the above-mentioned literature in the revised manuscript.

## 4 Response to Reviewer 4

**Comment:** More technical details for Adaptive Prediction Calibration should be provided, which serves as a major contribution of this paper.

Thank you for pointing this out. We have added more details for APC, detailed processes of which are described in the revised manuscript (Sec 4.2) and as follows.

Adaptive prediction calibration is the core component of our approach. In the hard class problem, target samples are prone to be classified into easy classes and difficult to be classified into hard ones, which leads to the lower predictive class proportion for hard classes. Meanwhile, One way to alleviate this problem is magnifying the predictive proportions of hard classes, keeping the normal ones, and suppressing the easy ones. To control predictive class proportions in a reasonable interval, we try to close the predictive class proportion and the true one of the target domain, which always is unknown in practice. Here, we replace it with the class proportion of the source domain.

The detail process of APC is shown in Figure 3 top. The target domain $\mathcal{D}_u$ is first fed into the trained model to obtain the predictions $P = \{p_i^u\}_{i=1}^{m_u}$. Then, we define a ratio $R$ of class distribution of source domain $q(y)$ to the predictive class distribution $p(y)$ as follows

$$R = q(y) \oslash p(y), \tag{1}$$

where $q(y) = \frac{1}{m_s}\sum_{i=1}^{m_s} y_i^s$, $p(y) = \frac{1}{m_u}\sum_{i=1}^{m_u} p_i^u$ and $\oslash$ means element-wise division, and $R$ is a $C$-dimensional vector with $i$-th dimension being the difficulty degree belonging to $i$-th class. Finally, we calibrate predictions $P$ by

$$P \leftarrow \{\text{Normalization}(R \odot p_i^u)\}_{i=1}^{m_u}, \tag{2}$$

where $\text{Normalization}(x) = \frac{x}{\sum_i x_i}$ and $\odot$ means element-wise multiplication. Intuitively, we calibrate $P$ by $R$. For a certain class $c$, if the predictive probability of class $c$ is small, which means that class $c$ is a hard class, the APC will increase the probabilities of classifying target samples into class $c$.

**Comment:** Since the objective of this paper is to improve the pseudo-labeling accuracy, more qualitative and quantitative results help show the superiority of the proposed algorithm.

Thank you for pointing this out. In the revised manuscript (Sec 5.6.2), we have reported more experimental results as follows.

We report the accuracy of pseudo-labels and learned network during training on Office-31 A→W and Office-Home Pr→Cl under UDA setting in Figure 5(a) and 5(b), respectively. We found that (1) As training processes, the test accuracy increases steadily, which illuminates the stability of our approach, and hence it can adapt to various scenarios better. (2) The test accuracy maintains a tight relationship with accuracy and the number of pseudo labels. In Office-31 A→W, the accuracy of pseudo labels keeps stable and the number increases, the test accuracy can keep step with the number of pseudo labels. Office-Home Pr→Cl is a quite challenging task, and our approach also can extract positive information and improve the test accuracy in the process of pseudo-labeling.

## References

[1] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring," *arXiv preprint arXiv:1911.09785*, 2019.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *CoRR*, vol. abs/1312.6199, 2014.

[3] Z. Cao, L. Ma, M. Long, and J. Wang, "Partial adversarial domain adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 135–150, 2018.

[4] P. Panareda Busto and J. Gall, "Open set domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 754–763, 2017.

[5] R. Shu, H. H. Bui, H. Narui, and S. Ermon, "A dirt-t approach to unsupervised domain adaptation," *ArXiv*, vol. abs/1802.08735, 2018.

[6] A. Kumar, P. Sattigeri, K. Wadhawan, L. Karlinsky, R. Feris, W. Freeman, and G. Wornell, "Co-regularized alignment for unsupervised domain adaptation," in *NeurIPS*, 2018.

[7] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3723–3732, 2018.

[8] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *International conference on machine learning*, pp. 5423–5432, PMLR, 2018.

[9] K. Saito, Y. Ushiku, and T. Harada, "Asymmetric tri-training for unsupervised domain adaptation," in *International Conference on Machine Learning*, pp. 2988–2997, PMLR, 2017.

[10] D.-D. Chen, Y. Wang, J. Yi, Z. Chen, and Z.-H. Zhou, "Joint semantic domain alignment and target classifier learning for unsupervised domain adaptation," *arXiv preprint arXiv:1906.04053*, 2019.

[11] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 5, pp. 1–46, 2020.

[12] J. Li, M. Jing, K. Lu, L. Zhu, and H. T. Shen, "Locality preserving joint transfer for domain adaptation," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 6103–6115, 2019.

[13] H. Wu, Y. Yan, G. Lin, M. Yang, M. Ng, and Q. Wu, "Iterative refinement for multi-source visual domain adaptation," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2020.

[14] S. Chen, M. Harandi, X. Jin, and X. Yang, "Domain adaptation by joint distribution invariant projections," *IEEE Transactions on Image Processing*, vol. 29, pp. 8264–8277, 2020.

[15] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *2013 IEEE International Conference on Computer Vision*, pp. 2960–2967, 2013.

[16] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 2066–2073, IEEE, 2012.

[17] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2010.

[18] M. Ghifary, D. Balduzzi, W. Kleijn, and M. Zhang, “Scatter component analysis: A unified framework for domain adaptation and domain generalization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1414–1430, 2017.

[19] L. Jing-jing, J. Mengmeng, L. Ke, Z. Lei, and S. Tao, “Locality preserving joint transfer for domain adaptation,” *arXiv: Computer Vision and Pattern Recognition*, 2019.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems* (P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1106–1114, 2012.

[21] M. Rahman, C. Fookes, M. Baktash, and S. Sridharan, “On minimum discrepancy estimation for deep domain adaptation,” *ArXiv*, vol. abs/1901.00282, 2020.

[22] H. Wu, Y. Yan, Y. Ye, M. Ng, and Q. Wu, “Geometric knowledge embedding for unsupervised domain adaptation,” *Knowl. Based Syst.*, vol. 191, p. 105155, 2020.

[23] J. Li, E. Chen, Z. Ding, L. Zhu, K. Lu, and H. T. Shen, “Maximum density divergence for domain adaptation,” *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[24] L. Hu, M. Kan, S. Shan, and X. Chen, “Unsupervised domain adaptation with hierarchical gradient synchronization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4043–4052, 2020.

[25] J. Yang, H. Zou, Y. Zhou, Z. Zeng, and L. Xie, “Mind the discriminability: Asymmetric adversarial domain adaptation,” in *European Conference on Computer Vision*, pp. 589–606, Springer, 2020.