Refining Pseudo Labels for Robust Test Time Adaptation

Huiwon Gwon Sunhee Jo Heajeong Jo Chanho Jung
Hanbat National University
Daejeon, South Korea

Abstract

Test-time adaptation(TTA) addresses challenges related to distribution shift by allowing the model to adapt to target data during testing without the need for source data. TTA has made significant advancements in the well-organized datasets sorted by distribution. However, it sometimes encounters challenges when applied to real-world data with diverse mixed distributions. In TTA, the pseudo labeling scheme performs properly on real-world data but significantly degrades when used with feature space information. Based on these observations, we propose a new pseudo-labeling scheme that does not rely on feature space information, bringing robustness in handling real-world data. Furthermore, Supervised Contrastive(SupCon) loss is leveraged to facilitate target feature learning guided by accurate pseudo labels. To adapt to diverse target distributions, we incorporate a 3x3 convolution layer parallel to the classifier block and train it. Our code is available at https://github.com/CVPR2024-VizWiz-Challenge-final

1. Introduction

Deep neural networks exhibit remarkable performance gains when the domains of the training and test data are the same. However, when the model experiences distribution shifts, it shows a significant performance drop. To address these challenges, test-time adaptation (TTA) has emerged, where the source data is no longer used to adapt to test data from a different distribution.

Existing TTA methods show significant performance improvements, especially in scenarios where the target domain remains static (e.g., ImageNet-C[4]). However, as shown in Tab. 1, these methods show somewhat similar performance to the source model when faced with real-world scenarios featuring mixed distributions. (e.g., VizWiz-Classification dataset[1]). Moreover, the pseudo labeling scheme performs well on real-world data but significantly degrades when used with feature space information. The reason for this is that, as demonstrated in previous research, different layers exhibit sensitivity to distinct types of domain shifts[3, 10], making them challenging to apply to real data with diverse distributions.

Method	Convnext-base
SOURCE	53.15 (+0.00)
MEMO[12]	53.30 (+0.15)
COTTAE10* [11]	53.35 (+0.20)
AdaContrast†[2]	50.05 (-3.10)
PLUE†[8]	44.80 (-8.35)
SHOT†[7]	52.50 (-0.65)
Ours	**60.50** (+7.35)

Table 1. Classification accuracy rate(%) for the VizWiz-Classification dataset. * indicates the sole application of the pseudo labeling scheme. † indicates the application of the pseudo labeling scheme with information from the feature space.

To avoid relying on feature space information, we introduce a teacher and student model structure, where the teacher model generates pseudo labels. Additionally, we propose a new strategy for refining pseudo labels, to ensure stable TTA even in real-world data. Specifically, pseudo-labels must be consistently outputted for each epoch in order to be utilized; otherwise, they are excluded. Moreover Augmentation-Averaged Pseudo Labels[9] and Pseudo-label uncertainty estimation[8] are applied additionally. Leveraging accurate pseudo labels as an alternative to GT labels enables the application of the Supervised contrastive (SupCon) loss[6]. We incorporate a parallel 3x3 convolution layer alongside the classifier block and train solely the 3x3 convolution layer, similar to [5].

2. Method

Overall framework. Our method consists of the teacher and student model. The teacher model outputs pseudo labels with the application of Augmentation-Averaged Pseudo Labels[9]. The student model incorporates a parallel 3x3 convolution layer alongside the classifier block and exclusively trains the 3x3 convolution layer, similar to [5].

Pseudo-labeling scheme. We introduce two methods. First, we examine all pseudo labels within a designated range of epochs, and if any inconsistencies are detected within this range, they are excluded. Second, we adopt the

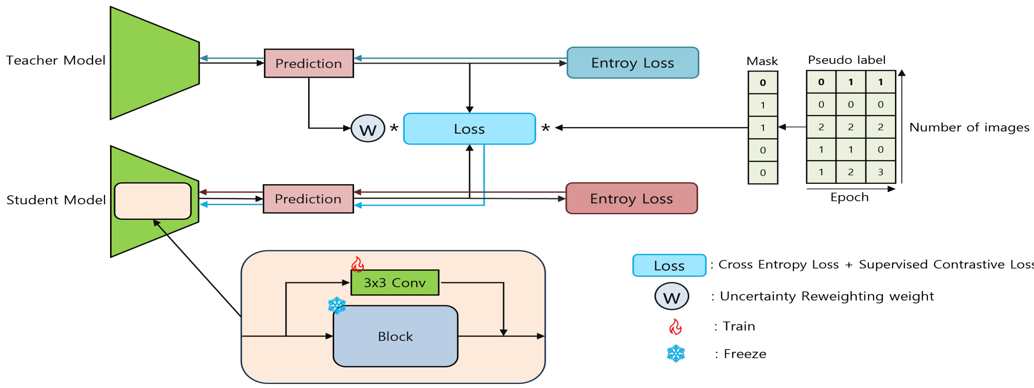

Figure 1. Overview of the proposed method.

Pseudo-label uncertainty estimation proposed in [8] to loss reweighting. Specifically, pseudo-labels from low entropy are considered to be more reliable, and more weight is assigned to the calculated loss of the corresponding pseudo-labels.

Diversity regularization. To prevent the model from overly relying on incorrect pseudo-labels during adaptation, we have adopted the approach suggested in [2].

Class mask. The prediction is filtered to calculate the loss only for those corresponding to the classes present in the dataset.

Supervised contrastive(SupCon) loss. With more accurate pseudo-labels, we are able to utilize the SupCon loss proposed in [6].

3. Experiment

Setup. In our experiments, we used the Convnext-base model with model weights sourced from the timm repository. We employed the SGD optimizer with a learning rate of 0.0025 and a batch size of 64.

Results. Tab. 2 shows ablation studies for individual components within the proposed method. First, when utilizing pseudo labels generated by the teacher model without relying on feature space information, performance significantly increased by +6.25% compared to the source model. Secondly, with the application of the pseudo-labeling scheme, accuracy increased by +0.3%, along with an additional performance boost of +0.2% facilitated by diversity regularization [2]. Furthermore, the integration of class masking resulted in a notable performance enhancement of +0.45%. Finally, employing the SupCon loss [2] with accurate pseudo labels led to a performance increase of +0.15%.

4. Conclusion

We introduce a novel pseudo-labeling scheme, enhancing TTA robustness in managing real-world data.

3x3 Conv layer	Modified Pseudo label	Modified Pseudo labeling Scheme	Diversity Regularization [2]	Class mask	SupCon Loss [2]	Total
						53.15
✓						51.40
✓	✓					59.40
✓	✓	✓				59.70
✓	✓	✓	✓			59.90
✓	✓	✓	✓	✓		60.35
✓	✓	✓	✓	✓	✓	**60.50**

Table 2. Ablation studies of individual components of the proposed method measured by classification accuracy on VizWiz Zero-Shot Classification.

References

[1] Reza Akbarian Bafghi and Danna Gurari. A new dataset based on images taken by blind people for testing the robustness of image classification models trained for imagenet categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16261–16270, 2023. 1

[2] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *CVPR*, 2022. 1, 2

[3] Gustavo A. Vargas Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Ismail Ben Ayed, and Christian Desrosiers. Clust3: Information invariant test-time training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6136–6145, 2023. 1

[4] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 1

[5] Xin Jin, Jia-Wen Xiao, Ling-Hao Han, Chunle Guo, Xialei Liu, Chongyi Li, and Ming-Ming Cheng. Make explict calibration implicit: "calibrate" denoiser instead of the noise model. 2023. 1

[6] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 1, 2

[7] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039, 2020. 1

[8] Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[9] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2022. 1

[10] Fahim Tajwar Ananya Kumar Huaxiu Yao Percy Liang Yoonho Lee, Annie S. Chen and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. In *Proceedings of the International Conference on Learning Representations*, 2023. 1

[11] Yongcan Yu, Lijun Sheng, Ran He, and Jian Liang. Benchmarking test-time adaptation against distribution shifts in image classification. *arXiv preprint arXiv:2307.03133*, 2023. 1

[12] M. Zhang, S. Levine, and C. Finn. Memo: Test time robustness via adaptation and augmentation. *arXiv preprint arXiv:2110.09506*, 2021. 1