Supplementary Material to "Memory-Efficient Pseudo-Labeling for Online Source-Free Universal Domain Adaptation using a Gaussian Mixture Model"

Pascal Schlachter, Simon Wagner, Bin Yang
University of Stuttgart, Germany
{pascal.schlachter, bin.yang}@iss.uni-stuttgart.de

A. Ablation studies

To gain deeper insights into our proposed GMM-based method, we conduct rich ablation studies in the following. Thereby, we use the OPDA scenario on the VisDA-C dataset as a representative example. The OPDA scenario effectively illustrates the challenging trade-off between rejecting new classes and reliably classifying a subset of known classes, while the VisDA-C dataset was selected arbitrarily.

A.1. Loss functions

Contribution of each loss: Fig. 1 illustrates the individual contributions of the contrastive loss $\mathcal{L}_\mathrm{C}$ and the KL divergence loss $\mathcal{L}_\mathrm{KLD}$ to the overall performance of our GMM-based method. Both losses prove to be effective and significantly enhance the source-only performance when used independently. Notably, the KL divergence loss thereby outperforms the contrastive loss, possibly because it impacts both the classifier and the feature extractor, whereas the contrastive loss only optimizes the feature extractor. However, the best results are achieved when combining both losses.

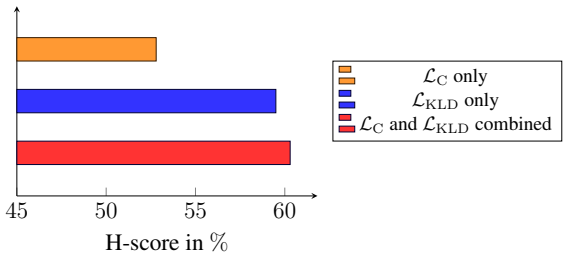

Figure 1. Results using different combinations of the losses $\mathcal{L}_\mathrm{C}$ and $\mathcal{L}_\mathrm{KLD}$ for the VisDA-C OPDA scenario.

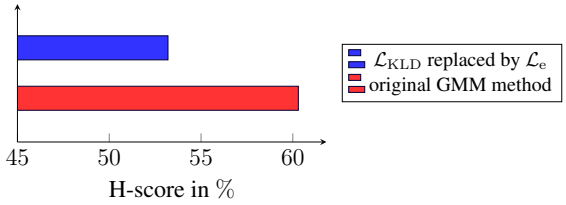

Figure 2. Results using COMET's entropy loss $\mathcal{L}_\mathrm{e}$ compared to the proposed KL divergence loss $\mathcal{L}_\mathrm{KLD}$ in our GMM method for the VisDA-C OPDA scenario.

Comparison between the KL divergence loss and the entropy loss of COMET: Our proposed KL divergence loss $\mathcal{L}_\mathrm{KLD}$ serves a similar function as the entropy loss $\mathcal{L}_\mathrm{e}$ used by COMET [10]. To evaluate their performance, we compare the original GMM method with a modified version where $\mathcal{L}_\mathrm{KLD}$ is replaced by COMET's entropy loss $\mathcal{L}_\mathrm{e}$. The results shown in Fig. 2 indicate that using COMET's entropy loss instead of the KL divergence loss leads to a significant decline in performance. This suggests that the KL divergence loss is more effective at encouraging confident predictions for known class samples while also guiding the model to appropriately handle OOD samples, making it the better choice for SF-UniDA.

A.2. Hyperparameter sensitivity

To analyze the sensitivity of our GMM-based method to hyperparameter choices, we vary each hyperparameter across a broad range around the chosen value while keeping the others constant. Figs. 3 to 7 show the results. Overall, our approach proves to be robust against these variations which indicates that the choice of hyperparameters is not critical for its success.

Number of dimensions of the reduced feature space: As shown in Fig. 3, stable performance is maintained when the number of dimensions FD_r of the reduced feature space is above 48. There is minor degradation at $FD_\mathrm{r} = 48$ and $FD_\mathrm{r} = 32$, respectively, before the performance drops significantly at $FD_\mathrm{r} = 16$. Therefore, the chosen number of feature dimensions $FD_\mathrm{r} = 64$ seems to provide the best trade-off between performance and memory efficiency.

Rejection rate: Fig. 4 shows that the performance remains nearly constant when varying p_reject between 25 %

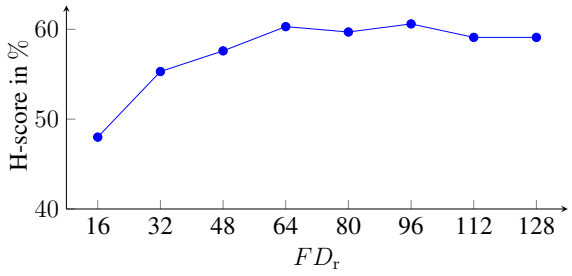

Figure 3. Results for the VisDA-C OPDA scenario using different numbers of dimensions for the reduced feature space.

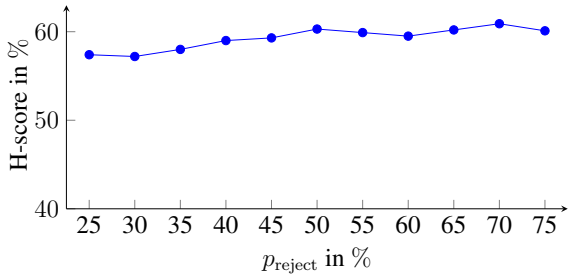

Figure 4. Results for the VisDA-C OPDA scenario using different rejection rates during pseudo-labeling.

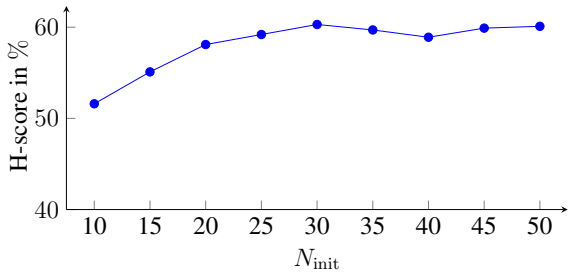

Figure 5. Results for the VisDA-C OPDA scenario using different numbers of batches for the initialization of τ_{k} and τ_{u} during pseudo-labeling.

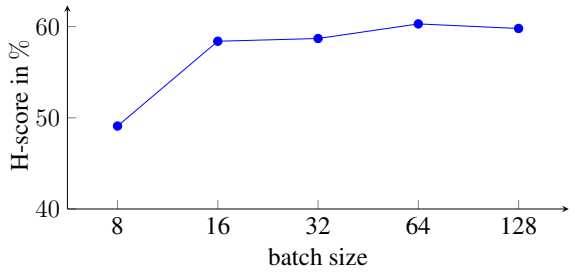

Figure 6. Results for the VisDA-C OPDA scenario using different batch sizes.

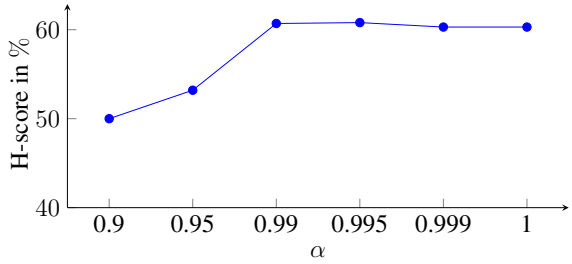

Figure 7. Results for the VisDA-C OPDA scenario using different exponential decay factors.

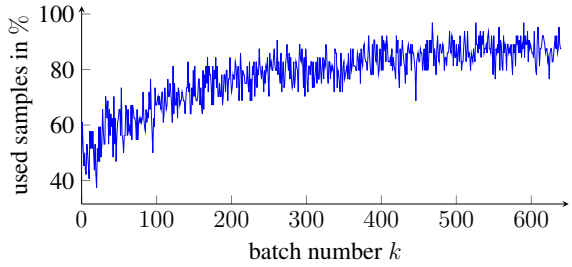

Figure 8. Development of the ratio of samples used for the adaptation over the course of an exemplary run of the VisDA-C OPDA scenario.

and 75 %. However, the performance slightly increases with a higher rejection rate. Therefore, discarding more samples during the initialization of τ_{k} and τ_{u} can be advantageous.

Number of initialization batches: Similar to FD_{r} and p_{reject}, the number of batches N_{init} used for initializing τ_{k} and τ_{u} does not significantly impact the performance of our method across a broad range of values, as shown in Fig. 5. However, a performance degradation is observed for values smaller than $N_{\mathrm{init}} = 20$. Thus, at least 20 batches are necessary for a representative initialization of τ_{k} and τ_{u}.

Batch size: In Fig. 6, we observe that decreasing the batch size starting from 128 only leads to a slight decline in performance until a batch size of 16. However, a significant performance drop occurs at a batch size of 8, which we suspect is due to the insufficient effectiveness of the initialization of τ_{k} and τ_{u} for such a small batch size.

Exponential decay factor: Fig. 7 shows that reducing the exponential decay factor from $\alpha = 1$ (no decay) results in a small improvement until $\alpha = 0.99$, after which performance declines sharply for $\alpha \leq 0.95$. Therefore, selecting $\alpha > 0.95$ appears to best balance the retention of valuable past information with the responsiveness to new data.

A.3. Ratio of samples used for adaptation

Fig. 8 illustrates how the percentage of samples used for adaptation evolves during the course of an exemplary run. For the first batch, $100\,\% - p_{\mathrm{reject}} = 50\,\%$ of samples are used. During the initialization of τ_{k} and τ_{u} in the first

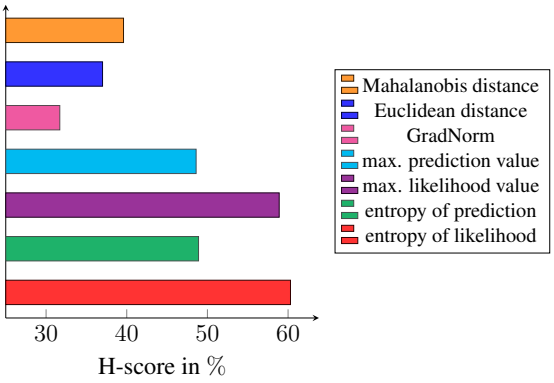

Figure 9. Results for the VisDA-C OPDA scenario using different metrics for the OOD detection.

$N_{\text{init}} = 30$ batches, this percentage remains relatively stable. Subsequently, it steadily increases on average, appearing to converge towards approx. $85\,\%$. This trend demonstrates the effectiveness of our adaptation method by showing that the pseudo-labeling becomes increasingly confident over time, leading to fewer samples being discarded due to uncertainty.

A.4. OOD metric

Besides the entropy of the likelihood $I(\boldsymbol{p}_{i,k})$, several other metrics can be used to evaluate the confidence of a sample belonging to the GMM distribution. These include the minimum Mahalanobis or Euclidean distance between the sample and the GMM means in the feature space [4], GradNorm [3], the maximum prediction value $\max_{c\in\mathcal{Y}_\text{s}} f_c(\boldsymbol{x}^\text{t}_{i,k})$, the maximum likelihood value $\max_{c\in\mathcal{Y}_\text{s}} p(\boldsymbol{x}^\text{t}_{i,k}|c;\hat{\boldsymbol{\mu}}_k(c),\hat{\boldsymbol{\Sigma}}_k(c))$, and the entropy of the prediction $I(f(\boldsymbol{x}^\text{t}_{i,k}))$. Fig. 9 presents a comparison of these OOD metrics for the VisDA-C OPDA scenario. It can be observed that the entropy of the likelihood provides the best result, while the maximum likelihood value performs only slightly worse. In contrast, all other evaluated OOD metrics yield significantly lower H-scores. Therefore, the two OOD metrics rooting on the likelihoods derived from the GMM prove to be the most robust, further demonstrating the effectiveness of our GMM-based knowledge transfer approach.

A.5. Vision transformer as backbone architecture

Tab. 1 presents the results on the VisDA-C dataset when the ResNet-50 backbone architecture is replaced with a Vision Transformer (ViT) [2], specifically the ViT-B/16 model. Its source training follows the same procedure as that used for the ResNet-50. Qualitatively, the results are similar to those obtained with the ResNet-50 backbone reported in the main paper. Notably, our proposed GMM-based method continues to achieve the best performance across all category shifts and therefore proves to also be effective for ViT-based model architectures.

Table 1. Results for the VisDA-C dataset using a Vision Transformer backbone. The accuracy (in %) is reported for PDA, while the H-score (in %) is reported for ODA and OPDA. Best results are in red, second best in blue.

	PDA	ODA	OPDA
Source-only	19.5	26.4	19.4
OWTTT [5]	31.3	54.1	49.9
COMET-P [10]	36.7	40.6	38.8
SHOT-O/P [6]	35.3	48.0	40.7
GLC [8]	9.9	3.4	9.6
GLC++ [9]	10.7	5.3	15.7
LEAD [7]	4.2	8.8	8.8
COMET-F [10]	31.3	39.5	38.6
GMM (Ours)	38.8	56.4	57.0

A.6. Measuring the absolute memory requirement

To validate the theoretical memory consumption analysis presented in section 3.7 of the main paper and provide additional insights, we measure the absolute memory required for the knowledge transfer across batches of COMET [10], a memory queue as proposed by [1], and our GMM-based approach. Additionally, we measure and compare the overall peak memory usage of COMET and our GMM-based method during adaptation, which is crucial for determining hardware requirements.

First, we consider the VisDA-C OPDA scenario where the number of source classes is $|\mathcal{Y}_\text{s}| = 9$. COMET requires a copy of the ResNet-50-based model to implement the student-teacher architecture, which consumes $94,098.23\,\text{KB}$ of memory. Second, in the memory queue, each prediction vector occupies $2.25\,\text{KB}$, and each feature vector requires $8.00\,\text{KB}$, resulting in a total of $10.25\,\text{KB}$ per sample. Finally our GMM-based method requires $4.50\,\text{KB}$ to store the means $\hat{\boldsymbol{\mu}}_k(c)$, $288.00\,\text{KB}$ for the covariance matrices $\hat{\boldsymbol{\Sigma}}_k(c)$ and $0.07\,\text{KB}$ for the weights $s_k(c)$, resulting in an overall memory consumption of $292.57\,\text{KB}$. Therefore, our GMM-based approach only requires

$$\frac{292.57\,\text{KB}}{94,098.23\,\text{KB}} \approx 0.0031 = 0.31\% \tag{1}$$

of memory compared to COMET. Moreover, starting with a memory queue size of only

$$\left\lceil \frac{292.57\,\text{KB}}{10.25\,\text{KB}} \right\rceil = 29$$

the memory queue already consumes more memory than our proposed GMM-based method.

Next, we consider an arbitrary OPDA scenario of DomainNet where the number of source classes is $|\mathcal{Y}_\mathrm{s}| = 200$. Due to the increased number of output neurons, the memory consumption of the ResNet-50-based model slightly increases to $94,290.73\,\mathrm{KB}$. Regarding the memory queue, each prediction vector now occupies $50.00\,\mathrm{KB}$ resulting in a total of $58.00\,\mathrm{KB}$ per sample. For the GMM-based method, the memory requirements increase to $100.00\,\mathrm{KB}$ for the means, $6,400.00\,\mathrm{KB}$ for the covariance matrices, and $1.56\,\mathrm{KB}$ for the weights, totaling $6,501.56\,\mathrm{KB}$. Hence, in this case, our GMM-based approach uses only

$$\frac{6,501.56\,\mathrm{KB}}{94,290.73\,\mathrm{KB}} \approx 0.0690 = 6.90\% \tag{2}$$

of the memory compared to COMET. Moreover, now starting from a queue size of

$$\left\lceil \frac{6,501.56\,\mathrm{KB}}{58.00\,\mathrm{KB}} \right\rceil = 113$$

the memory queue exceeds the memory usage of our GMM-based method. Although this value is considerably higher than for the VisDA-C dataset, it corresponds to significantly fewer than one sample per source class, which is still insufficient to enable reliable kNN- or clustering-based pseudo-labeling.

A similar picture is also observed when measuring the overall peak memory usage of COMET and our GMM-based method during each adaptation step. For the VisDA-C OPDA scenario, COMET's peak memory usage is nearly three times higher at $16,115.05,\mathrm{MB}$ compared to just $5,792.14,\mathrm{MB}$ for our GMM-based approach. Similarly, in the DomainNet OPDA scenario, COMET requires more than twice the memory, with a maximum of $16,113.88,\mathrm{MB}$ compared to $6,954.79,\mathrm{MB}$ for our method. Thus, as expected, the overall memory footprint of our GMM-based method is significantly smaller than that of COMET due to its more memory-efficient knowledge transfer. This reduced memory usage lowers hardware requirements, making our approach more suitable for real-world applications, particularly in embedded systems.

B. Discussions

B.1. Potential weaknesses of our method

Our method is based on the assumption that the target data classes form unimodal Gaussian-distributed clusters in the feature space. However, this may not always hold true. Although the contrastive loss encourages this clustering behavior, if the pseudo-labels are inaccurate at the beginning of the adaptation process due to a violation of this assumption, it can result in a slow initial adaptation. In the worst case, if the pseudo-labels are too inaccurate to trigger effective clustering, the adaptation may even fail completely. Although we did not encounter such a case in our experiments, to mitigate this risk, we recommend opting for a higher rather than lower initial ratio of left out samples p_reject.

A second potential failure case could occur with a highly imbalanced target dataset that contains only few samples of unknown classes. In this situation, since our KL divergence loss maximizes the divergence between the classifier output and a uniform distribution for samples pseudo-labeled as a known class, it may cause the classifier to converge to a trivial solution, i.e. always predicting the same class with high confidence. To prevent this from happening, in such a case, it might be advantageous to instead minimize the KL divergence between the prediction vector and the one-hot encoded pseudo-label for samples pseudo-labeled as a known class. By doing so, and maintaining a uniform pseudo-label for the unknown class, the KL divergence loss effectively becomes a cross-entropy loss. Nevertheless, also this case has never occurred during our experiments.

B.1.1 Difference between online and offline setting

When comparing the results from our online scenario with those achieved in offline SF-UniDA, like the results provided by GLC [8], GLC++ [9], and LEAD [7], a notable difference is evident. In the offline scenario significantly higher scores are achieved compared to online. For instance, the performance gap in the DomainNet OPDA scenario (considering only the domains painting, real, and sketch) is around $5\,\%$, while it extends to approximately $15\,\%$ in the VisDA-C OPDA scenario.

We believe that two main factors contribute to this discrepancy. First, offline methods have the advantage of being able to access all target data at once, which allows for more thorough adaptation strategies, such as kNN, which are not feasible in the online setting. Second, in offline scenarios, adaptation and prediction are separate steps, meaning the prediction only starts once the adaptation has been finished. This leads to an improved performance right from the start of the prediction. In contrast, in the online setting prediction and adaptation need to be performed in parallel which means that the prediction performance is initially equal to the source-only performance and only improves gradually. This difference in the adaptation process naturally reflects on the average performance scores.

References

[1] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 295–305, June 2022. 3

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at

scale. In *International Conference on Learning Representations*, 2021. 3

[3] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021. 3

[4] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 3

[5] Yushu Li, Xun Xu, Yongyi Su, and Kui Jia. On the robustness of open-world test-time training: Self-training with dynamic prototype expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11836–11846, 2023. 3

[6] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020. 3

[7] Sanqing Qu, Tianpei Zou, Lianghua He, Florian Röhrbein, Alois Knoll, Guang Chen, and Changjun Jiang. Lead: Learning decomposition for source-free universal domain adaptation, 2024. 3, 4

[8] Sanqing Qu, Tianpei Zou, Florian Roehrbein, Cewu Lu, Guang Chen, Dacheng Tao, and Changjun Jiang. Upcycling models under domain and category shift, 2023. 3, 4

[9] Sanqing Qu, Tianpei Zou, Florian Röhrbein, Cewu Lu, Guang Chen, Dacheng Tao, and Changjun Jiang. Glc++: Source-free universal domain adaptation through global-local clustering and contrastive affinity learning, 2024. 3, 4

[10] Pascal Schlachter and Bin Yang. Comet: Contrastive mean teacher for online source-free universal domain adaptation. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024. 1, 3