Analysis of Pseudo-Labeling for Online Source-Free Universal Domain Adaptation

Pascal Schlachter, Jonathan Fuss, Bin Yang
Institute of Signal Processing and System Theory, University of Stuttgart, Germany
{pascal.schlachter, bin.yang}@iss.uni-stuttgart.de

***Abstract*—A domain (distribution) shift between training and test data often hinders the real-world performance of deep neural networks, necessitating unsupervised domain adaptation (UDA) to bridge this gap. Online source-free UDA has emerged as a solution for practical scenarios where access to source data is restricted and target data is received as a continuous stream. However, the open-world nature of many real-world applications additionally introduces category shifts meaning that the source and target label spaces may differ. Online source-free universal domain adaptation (SF-UniDA) addresses this challenge. Existing methods mainly rely on self-training with pseudo-labels, yet the relationship between pseudo-labeling and adaptation outcomes has not been studied yet. To bridge this gap, we conduct a systematic analysis through controlled experiments with simulated pseudo-labeling, offering valuable insights into pseudo-labeling for online SF-UniDA. Our findings reveal a substantial gap between the current state-of-the-art and the upper bound of adaptation achieved with perfect pseudo-labeling. Moreover, we show that a contrastive loss enables effective adaptation even with moderate pseudo-label accuracy, while a cross-entropy (CE) loss, though less robust to pseudo-label errors, achieves superior results when pseudo-labeling approaches perfection. Lastly, our findings indicate that pseudo-label accuracy is in general more crucial than quantity, suggesting that prioritizing fewer but high-confidence pseudo-labels is beneficial. Overall, our study highlights the critical role of pseudo-labeling in (online) SF-UniDA and provides actionable insights to drive future advancements in the field. Our code is available at https://github.com/pascalschlachter/PLAnalysis.**

***Index Terms*—Pseudo-Labeling, Domain Adaptation, Universal, Online, Source-Free**

I. Introduction

Distribution shifts between training and test data often limit the real-world performance of deep neural networks. Accordingly, unsupervised domain adaptation (UDA) is necessary to adapt a pre-trained source model to unlabeled target data and, in this way, enables a robust performance across varying conditions. While early domain adaptation methods assume full access to source data, this is often impractical due to privacy concerns or memory constraints. Consequently, source-free UDA emerged, relying solely on the pre-trained source model and unlabeled target data for adaptation. This is often also referred to as test-time adaptation.

Considering the open-world nature of many real-world settings, recent advancements in source-free UDA aim to handle category (label) shifts alongside domain shifts. Specifically, this refers to scenarios where the source label space, $\mathcal{Y}_s$, and the target label space, $\mathcal{Y}_t$, are different ($\mathcal{Y}_s \neq \mathcal{Y}_t$). There are three cases: partial-set domain adaptation (PDA) ($\mathcal{Y}_t \subset \mathcal{Y}_s$), open-set domain adaptation (ODA) ($\mathcal{Y}_s \subset \mathcal{Y}_t$) and open-partial-set domain adaptation (OPDA) ($\mathcal{Y}_s \cap \mathcal{Y}_t \neq \emptyset$, $\mathcal{Y}_s \nsubseteq \mathcal{Y}_t$, $\mathcal{Y}_s \not\supseteq \mathcal{Y}_t$). Since prior knowledge of the category shift is typically unavailable, source-free universal domain adaptation (SF-UniDA) [1] aims to perform well across all three category shifts by detecting samples of new classes as unknown while accurately classifying samples of known classes.

Moreover, in many real-world applications, target data is received as a continuous stream rather than a fixed dataset with unrestricted access. To enable real-time processing of such data streams, adaptation and inference must be carried out simultaneously, with each test batch processed only once and handled sequentially. This setup is referred to as online UDA.

All existing methods towards SF-UniDA, both online [2], [3] and offline [1], [4], [5], rely on self-training with pseudo-labels. Thereby, the robustness of pseudo-labeling intuitively plays a crucial role in the success of the adaptation. However, the underlying relationship between pseudo-labeling and adaptation outcomes has not been studied yet. In this paper, we aim to address this gap and provide an extensive analysis of pseudo-labeling for online SF-UniDA. Specifically, we seek to answer the following research questions:

Q1. What is the upper bound of adaptation that would be achieved by a perfect pseudo-labeling and how far away is the current state-of-the-art from this upper bound?
Q2. What are the minimum requirements that pseudo-labeling must meet for adaptation to be effective, meaning that the performance improves w.r.t. the unadapted source model?
Q3. Existing SF-UniDA methods primarily rely on two types of loss functions: contrastive loss [2], [3] and cross-entropy (CE) loss [1], [4], [5]. How do they differ, particularly in terms of robustness to erroneous pseudo-labels?
Q4. Following [6], we exclude samples with uncertain pseudo-labels from the adaptation in our recent online SF-UniDA methods [2], [3]. The underlying intuition is that pseudo-label quality (accuracy) is more important than the quantity of samples used for adaptation. Does empirical evidence validate this hypothesis in practice?

II. Related Work

To date, the only universal domain adaptation (UniDA) methods that strictly adhere to the "source-free" definition given by [1]—meaning they rely solely on a standard pre-trained source model as the only source knowledge—are [1]–[5]. While other UniDA approaches [7]–[10] also claim

to be source-free because they do not access source data during adaptation, they impose specific training procedures or architectural constraints that differentiate them from strictly source-free methods. This limits their practical applicability.

Source-free UniDA (SF-UniDA) was first introduced by [1] with the Global and Local Clustering (GLC) method, which employs clustering-based pseudo-labeling. Adaptation is performed using a combination of CE loss and a kNN-based loss function. [4] extends this approach by incorporating an additional contrastive loss to enhance adaptation. Building on this foundation, [5] introduces Learning Decomposition (LEAD), a novel pseudo-labeling technique that separates features into source-known and -unknown components. This especially enables the use of individual rejection thresholds for unknown class identification during pseudo-labeling and allows the integration of confidence-weighting into the CE loss. Despite these variations, all these methods rely on entropy-based rejection for unknown classes during inference.

Recently, we demonstrated that these methods designed for offline SF-UniDA do not generalize well to the online setting because single batches fail to sufficiently capture the underlying data distribution [2], [3]. To address this limitation, we proposed two approaches specifically tailored for online SF-UniDA. Both methods follow similar adaptation strategies, combining a contrastive loss with an additional loss that enhances entropy-based distinguishability between known and unknown classes. However, they differ in their pseudo-labeling approaches: [2] employs a mean teacher, whereas [3] uses a Gaussian mixture model (GMM) in the feature space.

III. Simulation

A. Setup

1) Pseudo-Label Simulation: Pseudo-labeling is defined by two key metrics: pseudo-label accuracy and the proportion of samples used for adaptation, which we refer to as pseudo-label quality and quantity, respectively. To systematically investigate their impact on adaptation performance, we require a simulation that mimics real-world pseudo-labeling while enabling precise control over both factors. To achieve this, we first perform a forward pass without adaptation and compute the normalized entropy of the source model's predictions $\boldsymbol{p}_i$:

$$I(\boldsymbol{p}_i) = -\frac{1}{\log |\mathcal{Y}_s|} \cdot \boldsymbol{p}_i^T \cdot \log \boldsymbol{p}_i \tag{1}$$

where $|\mathcal{Y}_s|$ denotes the number of known classes.

To achieve a desired pseudo-label quantity by selecting samples for adaptation, we mimic the idea of [6], [2], and [3], which suggests that low-entropy samples indicate confident predictions of known classes, while high-entropy samples can be reliably classified as unknown. However, predictions with mid-range entropy values cannot be assigned to a class with confidence and therefore the corresponding samples are excluded from adaptation. To formalize this selection, we compute a confidence score based on the distance of each sample's entropy to the nearest extreme (i.e., 0 or 1):

$$d_i = \min(I(\boldsymbol{p}_i), 1 - I(\boldsymbol{p}_i)) \tag{2}$$

A certain pseudo-label quantity q is then achieved by selecting the top $q\%$ of samples with the smallest entropy distances d_i.

Next, to control the pseudo-label quality a, the selected samples are divided into two groups: those that receive correct pseudo-labels and those that are deliberately mislabeled. To maintain a realistic pseudo-labeling process, we again use the entropy distance d_i as an indicator. Samples with lower distances are more likely to be pseudo-labeled correctly while those with higher distances are more likely to be mislabeled. Specifically, among the selected samples, the top $a\%$ with the smallest distances d_i are assigned their correct pseudo-labels, while the remaining $(1-a)\%$ receive incorrect pseudo-labels.

Assigning correct pseudo-labels is straightforward, as the ground truth labels are provided by the dataset. However, assigning incorrect pseudo-labels requires careful consideration to ensure a realistic simulation of pseudo-labeling errors. Concretely, a realistic pseudo-labeling process reflects the model's natural tendency to confuse similar classes, meaning that misclassifications should follow the patterns observed in real predictions rather than being assigned arbitrarily. To achieve this, we assign incorrect pseudo-labels based on the model's predicted probabilities. If the true label corresponds to an unknown class, the sample is pseudo-labeled as the known class with the highest predicted probability. If the true label belongs to a known class, the sample is pseudo-labeled as the class with the highest predicted probability, excluding the correct label. However, if this probability falls below a threshold τ_i, the sample is instead pseudo-labeled as unknown.

To ensure that samples with higher prediction uncertainties are more likely to be pseudo-labeled as unknown, we define this probability threshold τ_i as an adaptive threshold that dynamically increases with entropy:

$$\tau_i = \alpha \cdot I(\boldsymbol{p}_i) \tag{3}$$

where α is a scaling factor and $I(\boldsymbol{p}_i)$ is the normalized entropy. Our analysis of a mean teacher [2] and GMM-based [3] pseudo-labeling revealed that when samples from known classes are incorrectly pseudo-labeled, they are predominantly assigned to the unknown class. To replicate this behavior, we empirically found that setting $\alpha = 1$, i.e. $\tau_i = I(\boldsymbol{p}_i)$, effectively mimics this tendency.

2) Loss Functions: Next, the generated pseudo-labels are used to guide the adaptation process. In the literature, SF-UniDA methods primarily employ two types of loss functions, which we aim to compare in this work.

The first approach combines a contrastive loss to encourage meaningful feature extraction with an entropy- or KL divergence-based loss to enforce separation between classifier outputs for known and unknown class samples [2], [3]. Here, we adopt the same loss function as COMET [2]. To maintain a strictly source-free setting, we avoid using source prototypes and instead compute target-domain prototypes using a class-wise exponential moving average based on the pseudo-labels.

The second approach primarily relies on a CE loss, where the unknown class is represented by uniformly distributed label vectors [1], [4], [5]. Although these works combine CE with

additional loss functions, we opt to use the CE loss alone in this study, as the other losses are either unsuitable for online settings—such as a kNN-based loss—or specifically designed to support particular pseudo-labeling strategies.

3) Dataset: For our experiments, we use the widely adopted public domain adaptation dataset VisDA-C [11], which covers a 12-class classification problem[1]. Its source domain contains 152,397 synthetically generated renderings of 3D objects while the target domain consists of 55,388 real-world images taken from the Microsoft COCO dataset [12]. To create category shift scenarios, we divide the 12 classes into shared classes (present in both domains), source-private classes, and target-private classes as follows:

- PDA: 6 shared, 6 source-private, 0 target-private
- ODA: 6 shared, 0 source-private, 6 target-private
- OPDA: 6 shared, 3 source-private, 3 target-private

4) Implementation details: We use the same experimental framework as [2]. The source model is based on a ResNet-50 [13] backbone and pre-trained using a standard supervised procedure. For domain adaptation, we use a batch size of 64 and optimize with SGD, applying a momentum of 0.9 and a learning rate of 0.001. During inference, we set a fixed entropy threshold of 0.5 to reject samples as unknown. Regarding the hyperparameters of the contrastive loss-based approach, we set the temperature to 0.1, the dimensions of the projection space to 128 and use $\lambda = 0.01$ as loss balancing factor.

We evaluate PDA using standard accuracy, while for ODA and OPDA, we report the H-score:

$$\text{H-score} = \frac{2 \cdot acc_k \cdot acc_u}{acc_k + acc_u} \tag{4}$$

which is the harmonic mean of the known and unknown class accuracies, acc_k and acc_u. Each experiment is repeated three times, and we report the average performance.

B. Results

The results are presented in the Tables I, II, and III. To enhance clarity and facilitate interpretation, we have color-coded each cell. Cells representing results worse than the source-only baseline are highlighted in red. For results improving upon the source-only baseline, a new color is introduced every 10%, with dark green shades finally indicating improvements greater than 40%. The source-only performance serves as the baseline, with an accuracy of 17.1% for PDA, an H-score of 31.7% for ODA, and an H-score of 26.9% for OPDA.

IV. Discussion

The experiments provide valuable insights into the relationship between pseudo-labeling and adaptation outcomes in online SF-UniDA. In the following, we discuss the key findings by answering the four formulated research questions.

[1]The classes are airplane, bicycle, bus, car, horse, knife, motorcycle, person, plant, skateboard, train, and truck.

A. Upper Bound of Adaptation (Q1)

As expected, the best results are achieved with perfect pseudo-labeling (100% pseudo-label quality and quantity). These results significantly surpass the current state-of-the-art achieved by [3], which yields an accuracy of 41.2% for PDA, an H-score of 59.9% for ODA, and an H-score of 60.3% for OPDA. Specifically, our simulation demonstrates that perfect pseudo-labeling achieves an accuracy of 75.6% for PDA, an H-score of 74.8% for ODA, and an H-score of 71.1% for OPDA using the contrastive loss-based approach. Even more strikingly, the CE loss-based approach reaches 91.9% for PDA, 76.7% for ODA, and 82.5% for OPDA with ideal pseudo-labeling. Hence, these results reveal a substantial gap between the upper bound of adaptation performance and the current state-of-the-art, highlighting the significant potential for improvement in existing methods. This margin underscores the critical role of pseudo-labeling accuracy and reliability in achieving optimal adaptation outcomes, suggesting that advancements in pseudo-labeling strategies could lead to substantial performance gains across all scenarios.

B. Pseudo-Label Requirements for Effective Adaptation (Q2)

The minimum pseudo-labeling requirements for effective adaptation depend heavily on the chosen loss function. Our results show that for the contrastive loss-based approach, adaptation remains effective even with a pseudo-label accuracy as low as 30%, as long as the pseudo-label quantity is sufficiently high ($\geq 40\%$). In contrast, when using CE loss, adaptation performance deteriorates much more rapidly as pseudo-label quality decreases, resulting in significantly higher minimum requirements. Specifically, a pseudo-label accuracy of at least 70% is necessary to achieve improvements over the source-only model across all category shift scenarios.

C. Comparison of Contrastive and Cross-Entropy Loss (Q3)

As already observed previously, the contrastive loss-based approach is significantly more robust to erroneous pseudo-labels while the CE loss is able to achieve substantially better results if the pseudo-label quality approaches 100%.

In a closer analysis, we find that the poor performance of the CE loss with low pseudo-label quality stems from the fact that it classifies nearly all samples as unknown. This behavior is not solely caused by the pseudo-labeling process misclassifying known classes as unknown, as setting $\alpha = 0$ (i.e., disallowing falsely unknown pseudo-labels) results in only small improvements. Instead, the issue appears to be inherent to the nature of the CE loss. We hypothesize that, due to its direct enforcement of desired model outputs, even small amounts of incorrect pseudo-labels can confuse the model, leading to many uncertain predictions that are subsequently interpreted as unknown.

In contrast, the contrastive loss operates on similarity constraints in the feature space. This allows to maintain better performance in cases where pseudo-labeling is imperfect, as it focuses more on relative relationships between samples rather than explicit class assignments. Consequently, the contrastive

TABLE I: Accuracies in % for the PDA scenario. The source-only baseline performance is 17.1%.

(a) Contrastive loss

pseudo-label quantity in % \ pseudo-label quality in %	0	10	20	30	40	50	60	70	80	90	100
10	1.9	8.0	10.4	7.9	28.5	20.8	34.0	48.1	49.2	57.7	63.7
20	0.9	1.6	9.9	22.5	27.6	24.9	45.4	44.9	51.2	54.2	71.8
30	0.7	1.5	12.0	21.5	23.4	28.2	37.4	39.8	41.3	54.6	74.3
40	7.0	5.0	10.4	17.5	23.9	25.8	35.0	38.5	43.5	53.1	74.6
50	2.9	7.6	11.8	22.3	24.4	19.7	29.6	37.3	45.4	38.5	74.4
60	2.0	6.6	7.0	19.1	19.9	21.0	26.9	26.8	29.0	36.1	73.6
70	3.8	6.9	9.6	18.6	21.3	19.4	26.8	25.2	31.6	44.5	75.0
80	1.6	11.0	11.5	19.8	18.3	26.4	24.4	27.2	26.8	35.9	74.7
90	4.6	8.3	10.8	17.2	18.5	17.7	21.6	22.7	29.5	41.2	74.8
100	4.4	6.2	17.8	17.9	18.1	19.0	18.4	23.4	29.0	33.1	75.6

(b) Cross-entropy loss

pseudo-label quantity in % \ pseudo-label quality in %	0	10	20	30	40	50	60	70	80	90	100
10	0.1	0.2	0.5	0.6	1.8	4.7	14.6	31.0	48.2	70.9	83.0
20	0.1	0.2	0.2	0.5	1.4	5.3	14.8	34.4	57.5	78.1	89.2
30	0.1	0.2	0.3	0.4	1.2	4.7	14.6	33.0	58.6	79.9	90.7
40	0.1	0.1	0.2	0.4	1.3	4.6	14.7	33.7	59.1	80.9	91.4
50	0.1	0.1	0.2	0.3	1.1	4.8	14.8	34.1	59.4	81.1	91.7
60	0.1	0.1	0.2	0.4	1.3	5.2	15.4	33.7	59.4	81.5	91.8
70	0.1	0.1	0.2	0.4	1.5	5.4	15.3	33.4	59.3	81.2	91.9
80	0.1	0.1	0.1	0.4	1.5	5.6	15.7	34.0	59.5	81.6	91.9
90	0.1	0.1	0.2	0.4	1.7	6.2	16.4	34.6	59.6	81.7	91.9
100	0.1	0.1	0.2	0.5	1.8	6.5	17.1	35.1	60.5	82.0	91.9

TABLE II: H-scores in % for the ODA scenario. The source-only baseline performance is 31.7%.

(a) Contrastive loss

pseudo-label quantity in % \ pseudo-label quality in %	0	10	20	30	40	50	60	70	80	90	100
10	4.9	9.1	10.0	12.0	19.4	28.5	40.6	46.9	50.8	53.3	54.7
20	10.9	8.1	14.9	26.8	33.8	45.1	50.8	57.1	59.1	60.0	63.7
30	9.8	15.2	23.9	28.5	37.5	47.9	54.2	59.0	60.4	62.7	64.8
40	15.6	21.3	28.9	33.6	44.5	48.4	53.0	58.5	63.0	61.6	68.1
50	20.4	23.8	30.6	38.1	45.0	53.4	56.8	60.4	64.0	65.3	70.0
60	22.1	25.3	30.6	37.9	46.5	52.0	59.2	61.6	65.6	68.1	71.1
70	20.7	27.4	36.8	41.8	49.3	52.0	59.6	61.1	65.2	69.2	73.3
80	23.2	27.5	35.9	42.0	49.4	54.2	60.0	63.5	67.0	69.7	73.3
90	21.5	35.0	33.3	41.5	51.1	55.8	58.6	63.0	67.0	69.7	75.1
100	22.3	27.1	34.6	45.5	51.1	56.3	62.6	64.1	67.7	71.8	75.2

(b) Cross-entropy loss

pseudo-label quantity in % \ pseudo-label quality in %	0	10	20	30	40	50	60	70	80	90	100
10	6.1	10.9	5.0	3.8	2.2	6.8	12.3	26.5	39.9	51.2	62.1
20	4.5	8.2	4.0	2.8	3.2	6.0	13.5	26.8	41.3	57.3	70.5
30	4.3	8.7	2.7	2.3	2.4	4.2	12.2	28.3	46.3	61.2	73.6
40	6.2	6.7	3.6	2.0	2.2	4.8	12.7	29.3	48.0	64.7	75.5
50	3.7	4.8	2.8	1.6	2.3	5.5	13.4	30.1	48.8	63.9	76.1
60	5.5	4.8	2.2	1.7	2.6	6.3	16.4	31.2	49.4	65.8	76.4
70	4.9	3.4	2.2	1.8	3.2	7.1	16.1	32.0	50.3	66.5	76.7
80	4.3	3.1	2.0	2.1	3.7	8.2	17.3	32.2	49.9	66.4	76.8
90	4.4	3.1	2.0	2.3	3.9	8.8	18.7	33.0	50.2	66.5	76.8
100	4.9	3.1	2.0	2.5	4.6	10.1	19.9	33.4	50.3	66.7	76.8

TABLE III: H-scores in % for the OPDA scenario. The source-only baseline performance is 26.9%.

(a) Contrastive loss

pseudo-label quantity in % \ pseudo-label quality in %	0	10	20	30	40	50	60	70	80	90	100
10	4.3	9.7	8.0	13.0	13.4	30.6	49.8	51.6	48.4	55.7	58.2
20	10.0	18.2	11.8	27.2	32.6	40.4	48.0	50.1	53.0	56.7	60.7
30	9.6	20.6	13.5	27.9	34.2	39.1	43.3	54.1	59.4	62.5	65.1
40	17.7	20.8	29.6	29.6	35.9	45.8	48.8	53.1	61.3	64.7	68.1
50	11.2	20.0	25.2	32.7	39.0	46.8	53.2	55.6	65.2	66.0	68.3
60	14.8	24.2	32.3	32.9	41.9	44.7	49.2	58.3	63.8	65.3	68.4
70	15.2	29.0	29.9	37.6	41.7	47.4	52.5	56.0	64.3	67.9	70.2
80	27.8	26.2	34.0	35.5	40.7	45.4	52.3	60.0	64.6	68.7	69.9
90	10.9	30.7	35.3	37.7	38.9	45.2	55.1	59.4	64.0	67.9	68.9
100	23.6	28.3	34.3	36.9	44.3	48.0	53.4	58.6	65.0	68.2	71.1

(b) Cross-entropy loss

pseudo-label quantity in % \ pseudo-label quality in %	0	10	20	30	40	50	60	70	80	90	100
10	0.4	1.7	1.4	2.0	4.5	13.7	24.8	41.2	55.6	67.7	76.1
20	5.2	3.9	1.5	2.1	4.9	13.0	26.0	44.4	60.2	73.8	80.0
30	3.4	1.2	1.3	2.0	5.0	12.7	27.5	44.5	63.1	75.1	81.8
40	2.8	3.4	1.8	2.0	5.9	14.4	28.0	45.4	62.9	75.6	82.1
50	2.9	1.9	2.2	3.1	8.3	17.3	29.0	46.2	63.8	76.3	82.3
60	2.3	2.4	2.3	4.3	10.8	18.4	29.7	46.4	63.3	76.6	82.5
70	3.1	2.4	3.1	5.6	12.2	19.3	31.1	46.2	64.4	76.9	82.3
80	2.9	2.3	2.9	7.0	13.1	20.7	31.4	47.2	64.3	76.6	82.4
90	2.9	1.9	3.0	8.6	13.9	21.3	31.7	46.9	64.2	76.6	82.3
100	2.5	1.4	2.9	8.9	14.8	21.5	32.0	46.4	63.7	76.5	82.5

loss provides a more stable adaptation process, making it the preferable choice when dealing with unreliable pseudo-labels.

D. Importance of Pseudo-Label Quality vs. Quantity (Q4)

The results confirm the intuition that pseudo-label quality plays a more critical role than quantity for effective adaptation. Across all scenarios, the adaptation performance generally declines much more rapidly as pseudo-label quality decreases compared to the impact of reduced pseudo-label quantity. This trend indicates that an adaptation approach focused on high-confidence pseudo-labels—despite using fewer samples—yields in general more effective results than one that relies on a larger number of uncertain pseudo-labels.

We assume that the reason for this observation is that incorrect pseudo-labels not only fail to improve the model but actively degrade its performance by steering adaptation in

the wrong direction. Intuitively, it is preferable to maintain the model unchanged rather than undergoing such negative adaptation. Accordingly, maximizing pseudo-label quality is the key priority, whereas quantity remains of secondary importance.

E. Further observations

In addition to addressing the core research questions, our analysis also reveals a noteworthy difference regarding the sensitivity to the pseudo-label quality between PDA and the other two category shifts. In the PDA scenario, which involves only known classes, adaptation performance is much more sensitive to pseudo-label quality. For example, using the contrastive loss-based approach, the performance difference between the best results for perfect pseudo-labeling quality (100%) and the best result for a pseudo-labeling quality of 90% is more than 17%, while this difference is below 4% for both ODA and OPDA. This discrepancy most likely arises because, in PDA, all predictions of samples as unknown directly correspond to prediction errors, as the target domain contains no unknown classes. In contrast, in ODA and OPDA, predicting samples as unknown is a necessary and desired behavior which needs to be encouraged by the adaptation as the source model is only trained on known classes. This fundamental difference highlights one of the main challenges of UniDA, where the adaptation method must perform well for all three scenarios without having prior knowledge. By following the tendency of the pseudo-labeling of [2] and [3] to predominantly misclassifying samples of known classes as unknown, this effect is even enhanced in the given simulation in favor of the ODA and OPDA scenarios. We could verify this statement by repeating one arbitrary experiment (quality 50%, quantity 100%) using $\alpha = 0$ for the pseudo-label simulation, meaning that no falsely unknown pseudo-labels occur. The result for PDA is with 41.4% more than 22% higher than the result of the original simulation. This insight emphasizes the need of SF-UniDA methods to balance the individual requirements of all three scenarios.

V. Limitations

Since our pseudo-label simulation was designed to analyze adaptation performance under a fixed, predefined global pseudo-label quality and quantity, it does not fully capture the dynamics of real pseudo-labeling during the adaptation process. In practice, adaptation itself influences pseudo-labeling, typically leading to an increase in batchwise pseudo-label quality and quantity over time.

In this sense, the choice of the loss function not only directly affects the adaptation outcome but also has an indirect influence by guiding the improvement of pseudo-labeling throughout the adaptation process. For instance, a contrastive loss promotes clustering of known classes in the feature space, which aligns well with GMM-based pseudo-labeling that assumes class distributions follow Gaussian clusters. However, our simulation, relying on synthetic pseudo-labeling, could not capture this interaction between the loss function and the evolution of pseudo-labeling. Therefore, in addition to the insights we presented in response to research question Q3, this effect should be considered when selecting a loss function and deserves further investigation.

VI. Conclusion

In conclusion, this work underscores the critical role of pseudo-labeling in online SF-UniDA and provides actionable insights for improving adaptation performance. By addressing the research questions, we contribute to a deeper understanding of the trade-offs between pseudo-label quality and quantity, the robustness of different loss functions, and the maximum achievable adaptation outcomes. The results highlight a significant gap between current state-of-the-art methods and the upper bound of performance achievable with perfect pseudo-labeling, emphasizing the need for further advancements in pseudo-labeling strategies. At the same time, the detailed insights and systematic analysis provided in this work offer valuable guidance for future research, paving the way for more effective approaches to narrow this gap and enhance adaptation performance in real-world applications.

References

[1] S. Qu, T. Zou, F. Röhrbein, C. Lu, G. Chen, D. Tao, and C. Jiang, "Upcycling models under domain and category shift," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 019–20 028.

[2] P. Schlachter and B. Yang, "Comet: Contrastive mean teacher for online source-free universal domain adaptation," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–9.

[3] P. Schlachter, S. Wagner, and B. Yang, "Memory-efficient pseudo-labeling for online source-free universal domain adaptation using a gaussian mixture model," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[4] S. Qu, T. Zou, F. Röhrbein, C. Lu, G. Chen, D. Tao, and C. Jiang, "Glc++: Source-free universal domain adaptation through global-local clustering and contrastive affinity learning," *arXiv preprint arXiv:2403.14410*, 2024.

[5] S. Qu, T. Zou, L. He, F. Röhrbein, A. Knoll, G. Chen, and C. Jiang, "Lead: Learning decomposition for source-free universal domain adaptation," 2024.

[6] Z. Feng, C. Xu, and D. Tao, "Open-set hypothesis transfer with semantic consistency," *IEEE Transactions on Image Processing*, vol. 30, pp. 6473–6484, 2021.

[7] J. N. Kundu, N. Venkat, R. V. Babu *et al.*, "Universal source-free domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4544–4553.

[8] J. N. Kundu, N. Venkat, A. Revanur, R. V. Babu *et al.*, "Towards inheritable models for open-set domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 376–12 385.

[9] J. Liang, D. Hu, J. Feng, and R. He, "Umad: Universal model adaptation under domain and category shift," 2021.

[10] X. Liu, Y. Zhou, T. Zhou, C.-M. Feng, and L. Shao, "Coca: Classifier-oriented calibration via textual prototype for source-free universal domain adaptation," 2024.

[11] X. Peng, B. Usman, N. Kaushik, D. Wang, J. Hoffman, and K. Saenko, "Visda: A synthetic-to-real benchmark for visual domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2021–2026.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.