A Unified Objective for Novel Class Discovery (Supplementary Material)

Enrico Fini[1] Enver Sangineto[1] Stéphane Lathuilière[2] Zhun Zhong[1*] Moin Nabi[3] Elisa Ricci[1,4]
[1] University of Trento, Trento, Italy [2] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[3] SAP AI Research, Berlin, Germany [4] Fondazione Bruno Kessler, Trento, Italy

1. Comparison with the state-of-the-art without self-supervised pretraining

In the main paper we presented a comparison with the state-of-the-art, using each method in its best possible configuration. For all competitors the best performance is reached by using self-supervised pretraining. This is disadvantageous for our method, since UNO only requires supervised pretraining. Also, using self-supervision makes all competing methods more computationally expensive.

Hence, in Tab. 1 we also report results without self-supervised pretraining, *i.e.* with supervised pretraining only. In this comparison, all methods are trained using roughly the same amount of compute. Clearly, all methods except UNO are negatively affected, *e.g.* RS loses ≈6%, DTC ≈10% on CIFAR100-20. As a consequence, UNO outperforms the competitors even more significantly (*e.g.* 6.7% and 17.6% on CIFAR10 and CIFAR100-20 respectively). This is a clear sign that, while UNO is able to learn powerful representations at discovery time, other methods need ad-hoc offline pretrainings that are often not possible in real world scenarios.

Method	CIFAR10	CIFAR100-20	ImageNet
k-means [8]	65.5±0.0	56.6±1.6	71.9
KCL [5]	66.5±3.9	14.3±1.3	73.8
MCL [6]	64.2±0.1	21.3±3.4	74.4
DTC [4]	87.5±0.3	56.7±1.2	78.3
RS [3]	89.4±1.4	67.4±2.0	82.5
UNO (avg)	**96.1±0.5**	**84.5±1.0**	**89.2**
UNO (best)	**96.1±0.5**	**85.0±0.6**	**90.6**

Table 1: Comparison with state-of-the-art methods on CIFAR10, CIFAR100-20 and ImageNet for novel class discovery using task-aware evaluation protocol. Clustering accuracy is reported on the unlabeled set (training split). All methods initialize the encoder with supervised learning on the labeled set. "RS+" is with incremental classifier.

*Corresponding author

2. Multi-view aggregation strategies

In Sec. 3 in the main manuscript, we described a way to generate pseudo-labels using multiple views. This corresponds to the **swapped prediction task** proposed in [1]. Nonetheless, other strategies can be employed. Instead of predicting the pseudo-label generated by another view, one can think of having a single pseudo-label that depends on all the views. This can be done in various ways. In the following we describe two aggregation strategies we investigated.

Averaging pseudo-labels. We generate pseudo-labels independently for each view using the Sinkhorn-Knopp algorithm [2]. Then, we aggregate the pseudo labels by simply averaging the pseudo label over the views. In the case of two views this is equivalent to computing the following:

$$\hat{\boldsymbol{y}} = \frac{\hat{\boldsymbol{y}}_1 + \hat{\boldsymbol{y}}_2}{2}. \tag{1}$$

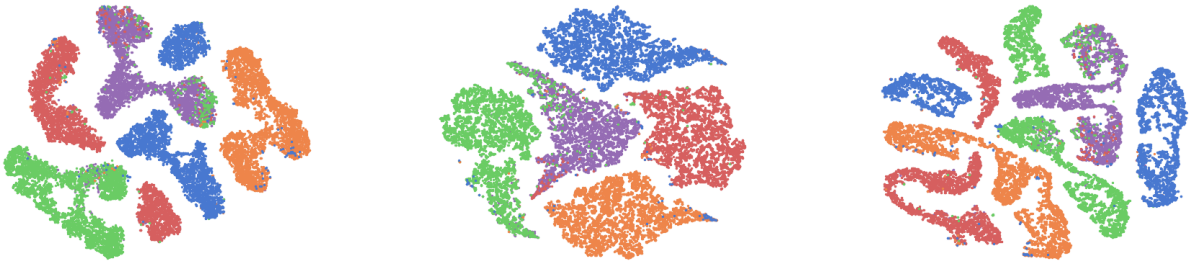
Subsequently, $\hat{\boldsymbol{y}}$ can be plugged in Eq. (2) of the main paper to obtain the complete pseudo-label.

This approach has advantages and disadvantages with respect to the swapped prediction task. For instance, an advantage is that the averaged pseudo-label is surely less noisy, since it depends on both views. However, averaging generates more entropic probability distributions, especially at the beginning, which slows down training. The parameters of the Sinkhorn-Knopp algorithm [2] can be tuned to account for the smoothing introduced by averaging. However, these parameters depend on the number of views, which makes this strategy impractical. Nonetheless, as shown in Tab. 2, this aggregation strategy produces very similar results to the swapped prediction task when using two views.

Averaging logits. Similarly, another solution to generate aggregated pseudo-labels is to first average the logits:

$$\boldsymbol{v}_g = \frac{\boldsymbol{v}_g^1 + \boldsymbol{v}_g^2}{2}, \tag{2}$$

then generate $\hat{\boldsymbol{y}}$ from $\boldsymbol{v}_g$, and finally plug Eq. (2) of the main paper. This solution is particularly advantageous in terms of



(a) Shared feature space (b) Clustering head logits (c) Overclustering head logits

Figure 1: t-SNE visualization of unlabeled samples (training set) on CIFAR10.

Method	Aggregation	CIFAR10	CIFAR100-50
UNO	logits	94.4	52.2
	pseudo-labels	**96.1**	52.5
	swap	**96.1**	**52.9**

Table 2: Comparison of multi-view aggregation strategies. "logits" stands for **averaging logits**, "Pseudo-labels" for **averaging pseudo-labels** and "swap" for the **swapped prediction task**. We report the clustering accuracy of the best head on the training set.

computation, since it requires us to run Sinkhorn-Knopp [2] only once, instead of twice (in the case of two views). However, unfortunately, this strategy does not produce results that are as good as using the swapped prediction task. (see Tab. 2)

3. More qualitative results for UNO

In this section, we show additional qualitative results and analysis of the feature space induced by our Unified Objective (UNO). In the main paper we reported a visual comparison of the features extracted by UNO w.r.t RS, showing a clear advantage of our method. Here, we dig deeper and investigate how clustering and overclustering heads project the features. For visualization purposes, we concatenate the logits of the multiple heads we use for clustering and overclustering respectively.

In Fig. 1a the reader can appreciate how unlabeled classes are organized in subgroups in the shared feature space. This is due to the use of overclustering. Nonetheless, these subgroups are tightly clustered and can be easily separated from samples of other classes. Indeed, as shown in Fig. 1b, the non linear projection head we make use of can correctly group most of the samples. Interestingly, the overclustering heads project the features in a way that increases the separation of the subgroups (see Fig. 1c), also sometimes stretching them in an attempt to minimize the loss.

Method	CIFAR100-20
DTC [4]	64.3
RS [3]	70.5
RS+ [3]	71.2
UNO (avg)	**74.7**
UNO (best)	**75.1**

Table 3: Comparison with state-of-the-art methods on CIFAR100-20 and ImageNet for novel class discovery with unknown number of classes C^u, using task-aware evaluation protocol. Clustering accuracy is reported on the unlabeled set (training split). All methods except UNO initialize the encoder with self-supervised learning.

4. Unknown number of clusters

All the results we showed so far assumed the knowledge of the number of classes C^u contained in the unlabeled set. However, in practical scenarios, it is unlikely to dispose of that information. While many previous works investigated the problem of estimating the number of clusters given a set of unlabeled data [8], in the context of NCD an effective approach was proposed in [4]. This method consists in holding out a probe subset from the labeled set, and then running a constrained (semi-supervised) k-means routine on the union of the probe subset and unlabeled set. Subsequently, the optimal number of clusters k is estimated using clustering quality indices on both subsets.

To investigate the applicability of our Unified Objective (UNO) in practical scenarios where k is not available, we

estimate the number of clusters on CIFAR100-20, using the aforementioned approach described in [4]. We use 60 classes for feature pretraining, 20 classes in the probe subset and 20 classes in the unlabeled set. In this way, we obtain a reasonable estimation, $k = 23$ classes. We then rerun UNO and the competing methods using this estimation. The results are shown in Tab 3. We find that our method still outperforms the state-of-the-art considerably.

Method	CIFAR10	CIFAR100-20	ImageNet
Jia *et al*. [7]	93.4±0.6	76.4±2.8	86.7
OpenMix [10]	95.3	-	85.7
NCL [9]	93.4±0.5	**86.6±0.4**	**90.7**
UNO (avg)	**96.1±0.5**	84.5±1.0	89.2
UNO (best)	**96.1±0.5**	85.0±0.6	90.6

Table 4: Comparison with concurrent methods on CIFAR10, CIFAR100-20 and ImageNet for novel class discovery using task-aware evaluation protocol. Clustering accuracy is reported on the unlabeled set (training split).

5. Comparison with concurrent works

In this section we compare the performance obtained with UNO to the following concurrent works: OpenMix [10], NCL [9] and Jia *et al*. [7]. The results can be found in Tab. 4. Despite being much simpler than all the concurrent related methods, UNO achieves better or comparable performance.

References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proc. NeurIPS*, 2020. 1

[2] Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proc. NeurIPS*, 2013. 1, 2

[3] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *Proc. ICLR*, 2020. 1, 2

[4] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proc. ICCV*, 2019. 1, 2, 3

[5] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *Proc. ICLR*, 2018. 1

[6] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *Proc. ICLR*, 2019. 1

[7] Xuhui Jia, Kai Han, Yukun Zhu, and Bradley Green. Joint representation learning and novel category discovery on single- and multi-modal data. In *Proc. ICCV*, 2021. 3

[8] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proc. BSMSP*, 1967. 1, 2

[9] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proc. CVPR*, 2021. 3

[10] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *Proc. CVPR*, 2021. 3