Supplementary material: Embedding contrastive unsupervised features to cluster in- and out-of-distribution noise in corrupted image datasets

1 Linear separability of samples on the hyper-sphere

Figure 1 illustrates the linear separability of samples on the hypersphere on for each class of CIFAR-10 corrupted with ID noise and OOD noise from ImageNet32. We train the unsupervised N-pairs algorithm and use a non-linear projection with the final dimension being 2 as in Wang *et al.* [10]. Here we use the simpler CIFAR-10 dataset because the final 2D projection size is too small, which causes convergence issues with more difficult classification problems such as CIFAR-100. The linear separation is not as good as when using the larger dimension of 128 for the contrastive projection head but allows direct visualization of the separation. We display $1,000$ randomly selected samples at the dataset level as well as the predicted linear boundary. The OOD samples cluster on one side of the circle, confirming our hypothesis by becoming linearly separable from the in-distribution data.

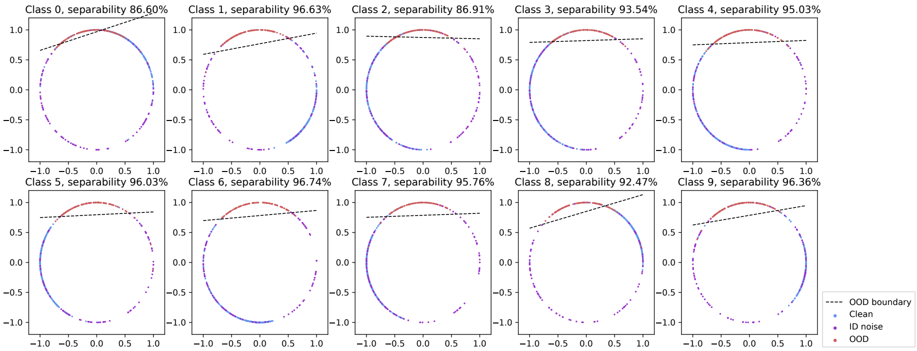

Fig. 1. Linear separation of OOD and ID on the hypersphere for each of the CIFAR-10 classes corrupted with $r_{in} = r_{out} = 0.2$, OOD from ImageNet32. Linear separability at the class level.

Table 1. Self-supervised initialization harms OOD robustness. CIFAR-100 corrupted with $r_{in} = r_{out} = 0.2$

Noise	Unsup init	CE	M	EDM	DSOS	RRL	Ours
$r_{in} = 0.2$	✗	65.79	67.50	71.03	70.54	72.64	72.95
$r_{out} = 0.2$	✓	64.89	66.99	56.89	66.76	66.76	71.62

Table 2. Linear separability between ID and OOD samples on the hypersphere in CIFAR-100 using ImageNet32 or Places365 as OOD data and on the miniImageNet part of CNWL dataset.

Dataset	Cifar-100		miniImageNet	
Corruption dataset	INet32	P365	Web	Web
r_{out}	0.2	0.2	0.2	0.6
Linear classifier score	98.21	95.95	99.66	99.54

2 Unsupervised initialization for label noise robust algorithms

In this paper, we chose to use the unsupervised features to detect the label noise but to avoid using them to initialize the CNN in the supervised phase. We did this to provide a fair comparison with existing noise robust algorithms as we observe that naive unsupervised initialization will harm the detection of out-of-distribution noise. Although existing works have showed that unsupervised learning can be used to improve robustness to in-distribution noise [2,3] or to reduce uncertainty [4] we find that the effects are detrimental in the presence of OOD noise. We believe that this is because unsupervised learning will learn features for OOD samples before the supervised phase, making it easier to overfit the OOD noise and reduce the capacity of existing algorithms to detect OOD images since all existing approaches rely on CNNs producing underconfident predictions on OOD samples [7,11]. Table 1 reports accuracy results of three state-of-the-art algorithms (EDM [9], DSOS [1], RRL [7]) and our algorithm trained using a CNN trained from random initialization or initialized using iMix [6] when trained on CIFAR-100 corrupted with 20% of ID and OOD noise from ImageNet32. We also report baselines that do not perform noise or label correction: CE and M [12]. We find that noise robust algorithms designed to function from a random initialization perform much worse when a self-supervised initialization is used. Note that part of the accuracy decrease could also be due to hyper-parameters that we did not tune. Standard cross-entropy training (CE) and Mixup (M) perform slightly worse as well or at least does not benefit from the unsupervised weight initialization. We encourage future research in this direction.

Table 3. hyper-parameters for training SNCF on corrupted image datasets.

Dataset	corruption	resolution	β	epochs	batch size	lr	lr red.	warmup	mixup	network
CIFAR-100	INet	32×32	1	100	256	0.1	$50, 80$	15	yes	PreActRes18
	Places	32×32	1	100	256	0.1	$50, 80$	30	yes	PreActRes18
CNWL	Web	32×32	$\tilde{r}_{out}$	100	256	0.1	$50, 80$	15	yes	PreActRes18
		299×299	1	200	64	0.01	$100, 160$	15	yes	InceptionResNetV2
Webvision	Web	227×227	1	100	64	0.01	$50, 80$	5	yes	InceptionResNetV2

3 Hyper-parameter table for experiments

Table 3 references the different hyper-parameters we use to train SNCF. We find that in most cases $\beta = 1$ is a sufficient hyper-parameter except for the CNWL dataset in the 32×32 resolution where we use the estimated ratio of OOD samples in the dataset $\tilde{r}_{out}$. This is not necessary for the 299×299 resolution where we use $\beta = 1$ for all noise configurations.

4 Run times

Although OPTICS is computationally expensive, we only run it once at the beginning of training on the embedded unsupervised features so it has no impact on the epoch train time. For reference, computation time for the spectral embedding + three OPTICS iterations for different neighborhood sizes $\{25, 50, 75\}$ on an i7-8700K in Python (averaged over 10 runs) takes 93s for 50,000 samples (100 classes) and 1h19m for 1,000,000 samples (1000 classes). This one-off cost is offset because we skip extra forward passes computed every epoch by other label robust algorithms (EDM, DSOS, SM) to evaluate feature representations or losses throughout training. Single epoch training times on a RTX 2080TI on CIFAR-100 with 20% OOD and ID noise are: ours 63s, EDM 93s, and DSOS 57s. Because of the equal sampling we perform in the algorithm, lower noise levels lead to longer run times as noisy examples are over-sampled to complete clean batches. This low noise configuration is the longest run time for SNCF. All algorithms are run using half precision.

5 Hardware

All networks are trained using mixed precision on a Nvidia RTX 2080TI (32×32 and 84×84 resolution) and two Nvidia TeslaV100 (227×227 and 299×299 resolution)

6 Visualization of OOD sample clustering on the CNWL

We observe that the samples can be clustered at the dataset level on the CNWL dataset [5] where little ID noise is present. Figure 2 is a UMAP [8] visualization of the separation of the features when the web noise is injected at 40% and a visual appreciation of how well the noise is captured by the 2D Gaussian mixture.

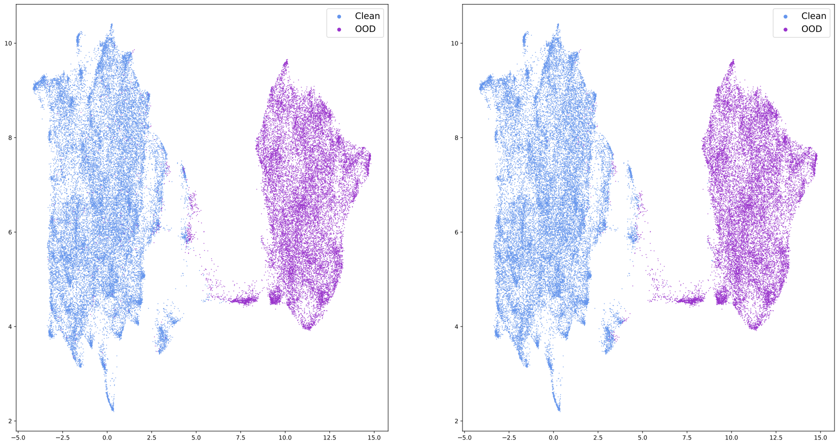

Fig. 2. Visualization of the separation of OOD clusters on the CNWL dataset corrupted with 40% web noise. The right hand plot shows the decision made by the 2D Gaussian Mixture when fit to the embedded features E while the left hand plot is the ground-truth.

7 Detected clusters of OOD samples on Webvision

Figure 3 shows examples of images from captured OOD clusters on Webvision. Some seem to capture very basic features such as black backgrounds in cluster 5, white backgrounds in cluster 10, or a basic square shape in cluster 1. More complex features are also captured such as human bodies in cluster 3, faces in cluster 4, hands in cluster 8, or text in cluster 9. Including these images when training the network helps improve low-level features and the accuracy when predicting on ID data (see the ablation table in the main paper).

8 Fast noise annotation using the linear separability of OOD samples

Although we propose to use an automatic algorithm to detect clusters in the ordered chain computed in OPTICS, we believe that for real world applications, this task could easily (and probably more accurately) be done by a human annotator, circling the clusters and selecting a few random images from each cluster to classify between clean or OOD at the cluster level. This would not be a lengthy task and would be favorable to real world applications.

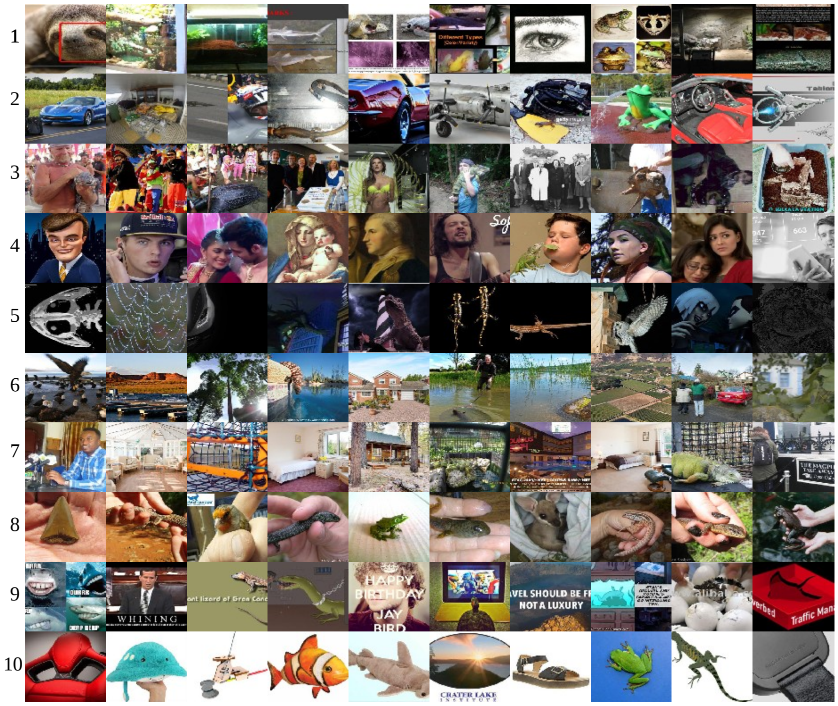

Fig. 3. Visualization of OOD clusters captured from the embedding.

Table 4. hyper-parameters for clustering the embedded contrastive features E using OPTICS.

Dataset	corruption	resolution	neigh. size	min size	ξ
CIFAR-100	INet	32×32	$\{75, 50, 25\}$	75	0.01
	Places	32×32	$\{75, 50, 25\}$	75	0.01
Webvision	Web	84×84	$\{75, 50, 25\}$	50	0.01

9 OPTICS hyper-parameters

Table 4 references the hyper-parameters used for detecting the OOD noise cluster and outliers, where ξ is the only sensitive hyper-parameter of the cluster detection algorithm in OPTICS that we provide a study for in the main body of the paper. We scan the embedded features at three neighborhood sizes and select the neighborhood leading to the lowest amount of outliers. For the CNWL dataset, we find that using a spherical covariance in the Gaussian mixture is better when the number of samples in the clean and OOD set is unbalanced, i.e. for 20% and 80% of noise, a spherical covariance captures the clusters more accurately. We use the full covariance setting otherwise.

10 SCNF algorithm

Algorithm 1 presents pseudocode for the SNCF algorithm.

Algorithm 1 SNCF

Input: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ a web noise dataset. h a randomly initialized CNN. g a CNN pretrained using self-supervised constrastive learning on $\mathcal{D}$

Parameters: $\alpha, e_{\text{warmup}}, e_{\max}, \xi, \tau$, neigh

Output: Trained neural network h

```
feats = $g(D)$                                          ▷ Extract unsup. contrastive features
embed = Embedding(feats, neigh)
for $c = 1, \ldots C$ do                                ▷ Apply OPTICS per class
    embedC = embed[class == c]
    clusterClean, clusterOod, outliers = OPTICS(embedC, $\xi$)
    allClean, allOod, allIdn ← clusterClean, clusterOod, outliers
end for

embedOod = Embedding(feats[allOod], neigh)              ▷ Re-embed without ID
simsOod = OPTICS(embedOod, $\xi$)                       ▷ Discover similar OOD clusters

for $e = 1, \ldots e_{\text{warmup}}$ do                ▷ Warmup
    for $t = 1, \ldots$ *numBatches* do
        Sample the next mini-batch $(x, y)$ from $\mathcal{D}[all_{\text{clean}}]$
        $l$ = CrossEntropy($h(x_{\text{mixed}}), y_{\text{mixed}}$)
        $h$ = UpdateNetworkWeights(L)
    end for
end for

for $e = e_{\text{warmup}} + 1, \ldots e_{\max}$ do     ▷ Noise robust training
    for $t = 1, \ldots$ *numBatches* do
        Two weakly augmented views $(x, y)$ and $(x', y)$ from $\mathcal{D}[allClean]$
        Two weakly augmented views $(z, )$ and $(z', )$ from $\mathcal{D}[allIdn]$
        $w$ = ConsistencyReg($h(z)$, $h(z')$, $\tau$)   ▷ Guessing labels for idn images
        Supervised mini-batch $X = (x, x', z)$ with labels $Y = (y, y, w)$
        $X_{\text{mix}}, Y_{\text{mix}}$ = mixup($X, Y, \alpha$)   ▷ Mixup
        $l_{\text{ce}}$ = CrossEntropy($h(X_{\text{mix}}), Y_{\text{mix}}$)

        Weak augs $Q = (x, z, o)$ from $\mathcal{D}[allClean], \mathcal{D}[allIdn], \mathcal{D}[allOod]$
        Strong augs $Q' = (x'', z'', o'')$ from $\mathcal{D}[allClean], \mathcal{D}[allIdn], \mathcal{D}[allOod]$
        sims = ComputeSims($y$, $w$, simsOod)
        $l_{\text{cont}}$ = GuidedContLoss($h(Q)$, $h(Q')$, sims)   ▷ Cont feats through proj.

        $h$ = UpdateNetworkWeights($l_{\text{ce}} + l_{\text{cont}}$)
    end for
end for
return $h$                                              ▷ Robustly trained network
```

References

1. Albert, P., Ortego, D., Arazo, E., O'Connor, N., McGuinness, K.: Addressing out-of-distribution label noise in webly-labelled data. In: Winter Conference on Applications of Computer Vision (WACV) (2022) 2
2. Cordeiro, F.R., Belagiannis, V., Reid, I., Carneiro, G.: PropMix: Hard Sample Filtering and Proportional MixUp for Learning with Noisy Labels. arXiv: 2110.11809 (2021) 2
3. Han, J., Luo, P., Wang, X.: Deep self-learning from noisy labels. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 2
4. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Using self-supervised learning can improve model robustness and uncertainty. In: Advances in neural information processing systems (NeurIPS) (2019) 2
5. Jiang, L., Huang, D., Liu, M., Yang, W.: Beyond Synthetic Noise: Deep Learning on Controlled Noisy Labels. In: International Conference on Machine Learning (ICML) (2020) 3
6. Lee, K., Zhu, Y., Sohn, K., Li, C.L., Shin, J., Lee, H.: i-Mix: A Strategy for Regularizing Contrastive Representation Learning. In: International Conference on Learning Representations (ICLR) (2021) 2
7. Li, J., Xiong, C., Hoi, S.C.: Learning from noisy data with robust representation learning. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 2
8. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv: 1802.03426 (2018) 3
9. Sachdeva, R., Cordeiro, F.R., Belagiannis, V., Reid, I., Carneiro, G.: EvidentialMix: Learning with Combined Open-set and Closed-set Noisy Labels. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2020) 2
10. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning (ICLR) (2020) 1
11. Yao, Y., Sun, Z., Zhang, C., Shen, F., Wu, Q., Zhang, J., Tang, Z.: Jo-SRC: A Contrastive Approach for Combating Noisy Labels. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2
12. Zhang, H., Cisse, M., Dauphin, Y., Lopez-Paz, D.: mixup: Beyond Empirical Risk Minimization. In: International Conference on Learning Representations (ICLR) (2018) 2