Appendix: ClusterMine: Robust Label-Free Visual Out-Of-Distribution Detection via Concept Mining from Text Corpora

Nikolas Adaloglou
Heinrich Heine University of Dusseldorf
adaloglo@hhu.de

Diana Petrusheva*
Heinrich Heine University of Dusseldorf
diana.petrusheva@hhu.de

Mohamed Asker*
Heinrich Heine University of Dusseldorf
dey59qad@hhu.de

Felix Michels
Heinrich Heine University of Dusseldorf
felix.michels@hhu.de

Markus Kollmann
Heinrich Heine University of Dusseldorf
markus.kollmann@hhu.de

A. Cluster-based metrics for ClusterMine.

To further investigate the quality of the clusters, we compute in Figure 1: (i) the intercluster **purity** (percent of samples withing a cluster that share the majority label), (ii) the intercluster **entropy** w.r.t. mined labels $\mathcal{Y}_{\text{pos}}$, and (iii) how often (percentage) $\mathcal{Y}_{\text{pos}}$ appear across multiple clusters (redundancy ratio), (iv) the ratio of mined $|\mathcal{Y}_{\text{pos}}|/C$. We find that clusters typically exhibit high purity relative to the number of clusters ($\geq 50\%$). This further supports our assumption that feature-space neighbors are label/cluster-consistent. Interestingly, the redundancy ratio increases as C increases, confirming that ClusterMine maintains semantic consistency while being robust to the overestimation of C. This analysis highlights the benefits of choosing ClusterMine over PosMine or NegLabel, where it is challenging to determine their respective hyperparameters in advance.

Heuristic for picking C. The redundancy ratio and the ratio of mined $|\mathcal{Y}_{\text{pos}}|/C$ could be used as a guideline to pick C using the elbow approach. While increasing C, these ratios tend to saturate and can serve as informative label-free heuristics in new application domains.

Text pre-processing. Homographs/duplicates words (e.g. bank) are deduplicated from the corpus, and we used only one lemma per Synset (no duplicates, one lemma in Tab. 1). In WordNet, a SynSet represents a group of cognitive synonyms that convey a shared concept or meaning. Nonetheless, text pre-processing had a minuscule impact on the reported results using ClusterMine as shown in Tab. 1.

*The authors contributed equally. Random order.

	No duplicates 1 lemma		Duplicates all lemmas		Duplicates 1 lemma	
	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95
NINCO	92.87	30.30	92.89	29.86	92.80	30.18
IN-O	93.57	29.40	93.53	28.60	93.38	30.45
OpenImage-O	96.93	15.91	97.06	14.90	96.85	16.09
iNat	99.00	4.77	99.05	4.14	98.92	4.98
IN-OOD	91.53	38.26	91.14	38.35	91.28	38.77
Textures43	93.45	32.89	94.10	27.61	93.99	29.82
Mean	94.56	25.25	94.63	23.91	94.54	25.05

Table 1. We report the impact of text-based preprocessing in WordNet (nouns and adjectives) using ClusterMine with CLIP ViT-H [3].

B. Results using additional OOD datasets.

Tab. 2 reported the AUROC on four additional OOD datasets. The Places dataset has the highest semantic overlap with the ID ($\approx 60\%$), while NINCOv2 has a near-zero semantic overlap, as it is a manually picked collection from existing OOD datasets. In contrast to prior works, we use Places as a bad benchmark to showcase how OOD detectors can reject samples that are more likely to be ID. We observe that MCM is the best-performing approach on Places and the worst-performing approach on NINCOv2, respectively. These two benchmarks could be utilized as OOD validation sets in future work.

Near-OOD and far-OOD detection. Recent works [2, 7] make a distinction between near-OOD and far-OOD based on image semantics or empirical difficulty. SSB, IN-OOD, and NINCO are considered near-OOD, while iNat, Textures, and OpenImage-O are considered far-OOD. Under this prism, ClusterMine is the current state-of-the-art

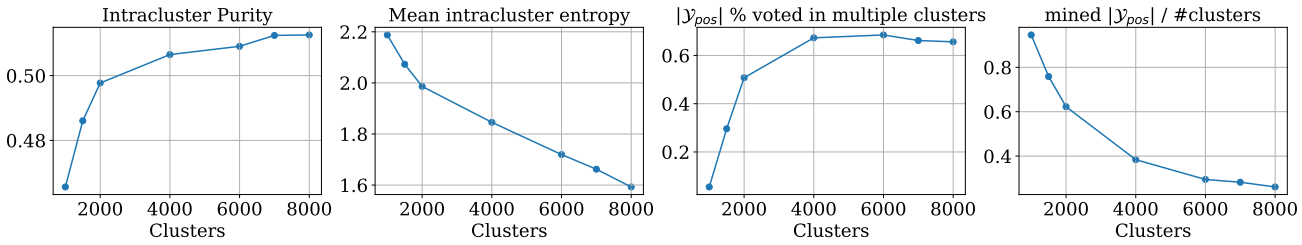

Figure 1. **ClustMine cluster analysis for various cluster sizes.** From left to right: intracluster purity measures the percentage of samples within a cluster that have the most frequent (majority) label (*left*), intracluster entropy is computed within samples in the same cluster (*center left*), and we compute the percentage of mined $|\mathcal{Y}_{\text{pos}}|$ that appears across multiple clusters as C increases (*center right*), and finally we compute the ratio of the mined $|\mathcal{Y}_{\text{pos}}|$ related to the chosen number of clusters C (*right*).

method on near-OOD detection.

C. Context prompt sensitivity.

In Fig. 2, we present the sensitivity of the label mining approaches to the different sets of context prompts. Basic refers to "An image of a {label}", while OpenAI refers to the set of 80 prompts as used by Radford et al. [6]. Ming refers to a subset of 5 prompts taken from the OpenAI set as used in [1, 5]. Nice refers to "A nice {label}" used by Jiang et al. [4]. Simple is a subset of 7 out of the 80 initial prompts, namely "itap of a {label}", "a bad photo of the {label}", "an origami {label}", "a photo of the large {label}", "a {label} in a video game", "art of the {label}", "a photo of the small {label}", based on follow-up analysis of Radford et al. [6] https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb. We adopt the simple prompts for all the reported results in the main text.

D. Additional results using negative label mining.

Negative label mining on POS and negative grouping strategy. Fig. 3 shows that negative label mining is not improving performance even for a large-scale text corpus such as POS. We included all available nouns and adjectives. In Fig. 4, we show that the proposed negative grouping strategy by NegLabel [4] deteriorates the OOD detection AUROC. Thus, we did not include it in our experiments.

Negative mining on various covariate ID sets. In Fig. 5, we demonstrate that negative label mining is not the reason of superior performance on the datasets with stylistic perturbations, namely ImageNet-R and Sketches. Hence, the outperformance of NegLabel is primarily attributed to the *a priori* knowledge of $\mathcal{Y}_{\text{GT}}$.

Table 2. **Semantic OOD detection detection AUROC on additional OOD datasets.**

Method	Places	Texture	SUN	SSB	NINCOv2
MCM	**92.32**	90.40	94.37	79.69	92.50
PLP	92.15	**93.08**	94.01	81.42	94.72
NegLabel	90.08	83.13	93.70	82.80	93.07
NegLabel*	90.39	88.29	94.41	85.13	94.53
PosMine	92.08	92.45	**95.51**	85.44	95.58
ClusterMine	91.41	91.80	94.70	**86.04**	**95.86**

E. Combining pseudo-label-probing (PLP) with positive label mining.

Finally, we explore whether pseudo-label probing (PLP) can be combined with positive label mining in Appendix D. We found no significant performance gain by combining PLP with ClusterMine and PosMine.

F. Robustness to covariate ID shifts.

Table 4 shows all the individual robustness scores for the sensitivity to ID shifts. In the main text we report the mean AUROC and FPR95.

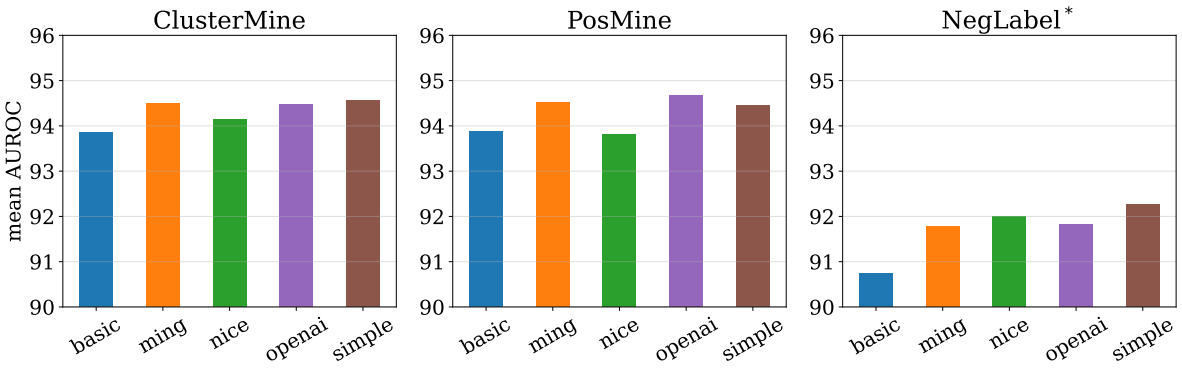

Figure 2. **Ablation study on context text prompts using CLIP VIT-H dfn5b [3].**

Table 3. **Applying pseudo-label probing (PLP) using the derived $\mathcal{Y}_{\text{pos}}$ instead of $\mathcal{Y}_{\text{GT}}$ from PosMine and ClusterMine, similar to [1].**

Method	$\mathcal{Y}_{\text{GT}}$ /$\mathcal{Y}_{\text{corpus}}$	NINCO	IN-O	OpenImage-O	iNat	IN-OOD	Textures43	Average AUROC / FPR95
PLP [1]	✓/✗	91.80	93.30	**97.87**	98.94	90.01	94.79	94.45 / 27.29
PosMine	✗/✓	92.56	93.13	97.04	98.83	91.36	93.81	94.46 / 26.41
PosMine + PLP	✗/✓	91.72	92.79	97.43	98.68	88.56	94.38	93.93 / 27.79
ClusterMine	✗/✓	**92.87**	**93.57**	96.93	**99.00**	**91.53**	93.45	**94.56 / 25.26**
ClusterMine + PLP	✗/✓	92.25	93.07	97.46	98.79	88.89	94.71	94.20 / 27.53

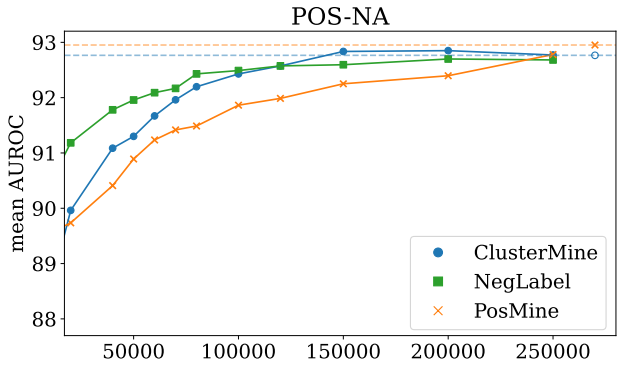

Figure 3. **The impact of negative label mining on the POS-NA corpus.**

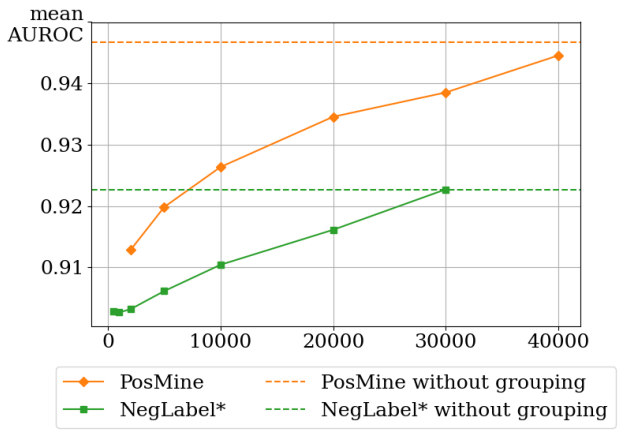

Figure 4. **Ablation study on the negative grouping strategy using CLIP VIT-H dfn5b [3].** The x-axis represents group size (i.e. number of label names per group). NegLabel* uses 40K negative mined labels as in the main manuscript.

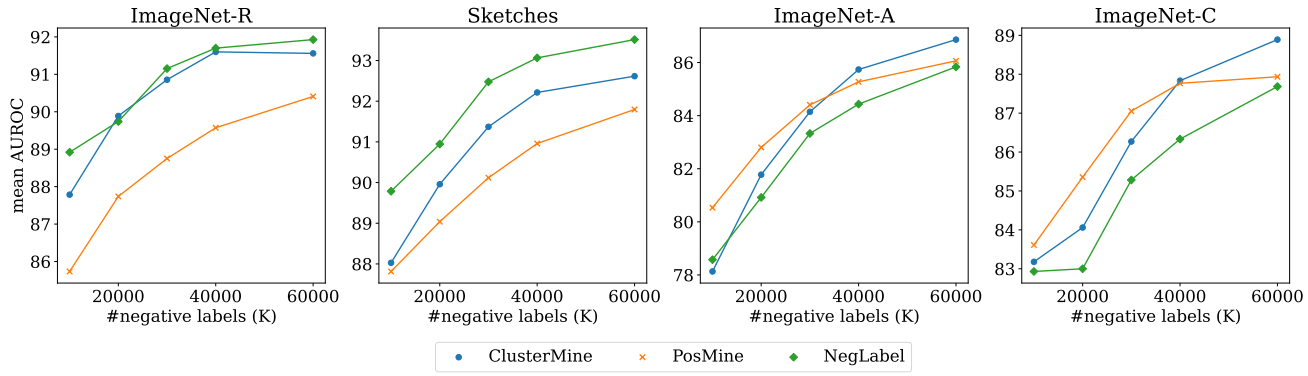

Figure 5. **OOD detection robustness to covariate ID shifts.** The superior performance of NegLabel on ImageNet-R and Sketches is not attributed to the negative label mining but rather to the a priori knowledge of $\mathcal{Y}_{\text{GT}}$, unlike PosMine and ClusterMine .

Table 4. **Quantifying robustness to in-distribution shifts using CLIP ViT-H dfn5b [3].** We report AUROC (%,↑) per OOD detection benchmark as well as mean AUROC and mean FPR95 (%,↓). Results with MCM [5] and NegLabel [4], where NegLabel* uses 40K negative mined labels. The highlighted light gray values highlight the discussion point raised in the main manuscript.

ID Dataset	Method	NINCO	IN-O	OpenImage-O	iNat	IN-OOD	Textures43	Average AUROC	Average FPR95
ImageNet-Sketches	MCM	84.88	87.73	94.18	94.01	85.77	88.22	89.13	55.74
	NegLabel*	91.21	91.11	96.31	98.85	89.50	91.41	**93.06**	40.29
	PosMine	89.60	90.44	95.68	98.24	88.14	91.25	92.22	32.92
	ClusterMine	90.35	91.14	95.60	98.52	88.43	90.76	92.47	**30.94**
ImageNet-R	MCM	77.67	81.64	91.08	90.74	78.77	82.11	83.66	68.28
	NegLabel*	89.30	89.48	95.44	98.59	87.71	89.68	**91.70**	41.42
	PosMine	87.74	88.69	94.91	98.00	86.01	89.58	90.82	**35.92**
	ClusterMine	88.34	89.13	94.50	98.17	85.91	88.61	90.78	37.50
ImageNet-A	MCM	62.95	68.79	84.49	83.75	64.15	69.05	72.20	76.80
	NegLabel*	79.70	80.82	90.74	96.64	78.14	80.56	84.43	56.32
	PosMine	82.13	83.52	92.19	96.71	79.92	84.64	86.52	44.44
	ClusterMine	83.53	84.45	91.88	97.17	80.04	83.46	**86.76**	**42.57**
ImageNetV2	MCM	84.72	87.79	94.79	94.66	85.66	88.21	89.31	48.97
	NegLabel*	87.91	87.97	94.60	98.17	86.09	88.22	90.49	45.38
	PosMine	90.44	91.10	95.94	98.31	89.00	91.86	**92.78**	32.15
	ClusterMine	90.70	91.39	95.65	98.49	88.85	91.10	92.70	**31.87**
ImageNet-C	MCM	70.01	73.90	84.58	83.76	70.76	74.43	76.24	89.99
	NegLabel*	82.68	83.18	91.54	96.49	80.95	83.17	86.33	54.87
	PosMine	84.86	85.61	92.60	96.47	82.80	86.58	88.15	52.10
	ClusterMine	86.40	87.02	92.81	97.11	83.75	86.54	**88.94**	**49.21**
ImageNet ID test set	MCM	88.78	91.30	96.64	96.62	89.65	91.75	92.46	35.22
	NegLabel*	90.26	90.10	95.79	98.64	88.40	90.44	92.27	40.43
	PosMine	92.56	93.13	97.04	98.83	91.36	93.81	94.46	24.78
	ClusterMine	92.87	93.57	96.93	99.00	91.53	93.45	**94.56**	**24.23**

References

[1] Nikolas Adaloglou, Felix Michels, Tim Kaiser, and Markus Kollmann. Adapting contrastive language-image pretrained (clip) models for out-of-distribution detection. *arXiv e-prints*, pages arXiv–2303, 2023. 2, 3

[2] Faruk Ahmed and Aaron Courville. Detecting semantic anomalies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3154–3162, 2020. 1

[3] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 1, 3, 4

[4] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided OOD detection with pretrained vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 4

[5] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in Neural Information Processing Systems*, 35:35087–35102, 2022. 2, 4

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[7] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023. 1