NegRefine: Refining Negative Label-Based Zero-Shot OOD Detection

Supplementary Material

A. Extended Experiments

A.1. OOD Dataset Descriptions

We used the following OOD datasets for ImageNet-1K:

- iNaturalist [34]: iNaturalist is a dataset containing over 5,000 species of plants and animals. In [15], a subset of 110 plant classes not present in ImageNet-1K was manually selected, and 10,000 images were randomly sampled from these classes to serve as OOD for ImageNet-1K.
- OpenImage-O [37]: This dataset consists of 17,632 samples from the OpenImage-v3 test set, which were manually verified as OOD with respect to ImageNet-1K.
- NINCO [3]: NINCO consists of 5,879 samples from 64 classes, collected from various sources, including existing datasets and the internet. Each sample was manually verified to be truly OOD with respect to ImageNet-1K.
- Clean [3]: Clean is a collection of 2,715 OOD images obtained from an analysis of 400 random samples drawn from 12 common OOD datasets (such as PLACES, Textures, Species, SSB-HARD, etc.), which were manually evaluated to determine whether they were truly OOD.

A.2. Baselines Implementation Details

For all baselines, we used the original source code from their GitHub repositories. The only exception is ZOC, where, following [16, 22], we upgraded the caption generator to BLIP (a more advanced model) to improve its effectiveness on ImageNet-1K; other implementation details for this method remain unchanged from the original paper.

A.3. Randomness and Standard Deviations

CSP involves randomness in generating modified conjugated class labels, where adjectives from WordNet are concatenated with a randomly selected term from a predefined list of 14 superclass labels. Our method builds on CSP and inherits this same source of randomness. To account for this, we repeated the experiments for both our method and CSP 10 times (using seeds 0 to 9). Additionally, CLIPN requires training an extra model on top of CLIP, and the authors have provided model snapshots from three runs in their GitHub repository. Other baselines do not involve any randomness. Tab. 6 reports the average and standard deviation for our method and CSP (based on the ten repetitions), as well as for CLIPN (based on the three model snapshots).

A.4. Extended Analysis and Ablations

Ablation on the Size of the Initial Negative Label Set. For the initial set of negative labels, we relied on CSP, which, similar to NegLabel, selects the top $p = 15\%$ of words least related to the in-distribution labels as negative labels. Tab. 7 presents an ablation study evaluating different values of p and comparing the performance of our method with CSP and NegLabel. Our method consistently outperforms both methods in average AUROC and FPR95 across all values of p. The lowest average FPR95 for our method is achieved at $p = 40\%$, while CSP and NegLabel reach their optimal FPR95 at $p = 60\%$ and $p = 30\%$, respectively.

Ablation on Different LLMs. Our negative label filtering mechanism (NegFilter) leverages an LLM to identify proper nouns and subcategories. Tab. 8 reports an ablation study on four mid-sized open-source LLMs. The results show that all LLMs yield nearly identical performance, suggesting that our method is not sensitive to the LLM choice.

Ablation on Different CLIP Architectures. In the main paper, following the literature, we adopted the CLIP ViT-B/16 model for our method and all baselines. In Tab. 9, we compare the performance of our method with CSP and NegLabel across different CLIP architectures. The results show that our method consistently outperforms both methods across all architectures by a significant margin.

Ablation on the Design and Parameters of S_{MM}. In Tab. 10, we present an ablation study on alternative designs for the S_{MM} score function and different values of the α parameter. We also report an ablation on varying k in Tab. 11. Notably, performance changes significantly from $k = 1$ to $k = 2$, while other k values yield almost similar results.

A.5. Evaluation Using Additional In-Distribution Datasets

Following NegLabel and CSP, we also evaluate our method using other in-distribution datasets, including Stanford-Cars [17], CUB-200 [35], Oxford-Pet [29], and Food-101 [4], considering iNaturalist [34], SUN [40], Places [44], and Textures [7] as OOD. Tab. 12 compares our method with CSP and NegLabel. All methods achieve near-optimal performance across most datasets, with CSP and NegLabel slightly outperforming ours in some cases. Notably, these in-distribution datasets belong to narrow and distinct domains (cars, birds, pets, and food), making them easily distinguishable from the selected OOD datasets. As a result, all three methods exhibit similar performance.

Methods	iNaturalist		OpenImage-O		Clean		NINCO	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
CLIPN	96.20 ± 0.41	19.11 ± 2.07	92.22 ± 0.42	30.36 ± 1.59	87.31 ± 0.46	41.56 ± 1.71	78.72 ± 1.19	66.51 ± 1.82
CSP	**99.60** ± 0.00	1.54 ± 0.02	94.09 ± 0.03	28.94 ± 0.12	88.32 ± 0.05	38.75 ± 0.17	77.88 ± 0.17	68.65 ± 0.18
NegRefine (ours)	99.57 ± 0.00	**1.47** ± 0.03	**95.00** ± 0.06	**23.85** ± 0.16	**90.65** ± 0.12	**33.18** ± 0.37	**81.92** ± 0.26	**61.96** ± 0.41

Table 6. Results including standard deviations for methods that involve randomness, corresponding to Table 1 in the main paper. For our method (NegRefine) and CSP, results are averaged over 10 different seeds. CLIPN results are based on the three model snapshots available in the authors' GitHub repository. Other baselines did not involve randomness. See Sec. A.3 for more details.

Methods	p%	iNaturalist		OpenImage-O		Clean		NINCO		Average	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
NegLabel	15	99.52	1.85	93.74	28.62	86.79	41.22	77.30	68.70	89.34	35.10
	20	99.41	2.30	93.95	27.83	86.72	40.85	76.78	68.32	89.22	34.82
	30	99.20	3.16	94.17	27.25	86.67	40.55	76.61	67.78	89.16	34.68
	40	99.01	3.90	94.23	27.21	86.46	41.18	75.96	67.78	88.91	35.02
	50	98.84	4.71	94.23	27.11	86.28	41.51	75.54	67.81	88.72	35.29
	60	98.67	5.53	94.20	27.28	86.04	41.88	75.19	67.74	88.53	35.61
	70	98.50	6.36	94.15	27.43	85.80	42.69	74.75	67.95	88.30	36.11
	80	98.34	7.17	94.08	27.50	85.58	43.09	74.45	67.98	88.11	36.44
	90	98.19	7.84	94.01	27.86	85.37	43.79	74.10	68.65	87.92	37.04
	100	98.05	8.53	93.95	28.05	85.17	44.42	73.86	68.59	87.76	37.40
CSP	15	99.60	1.54	94.09	28.94	88.32	38.75	77.88	68.65	89.97	34.47
	20	99.54	1.72	94.36	28.11	88.31	38.42	77.41	68.53	89.90	34.19
	30	99.39	2.36	94.66	26.68	88.22	38.08	77.64	67.73	89.98	33.71
	40	99.27	2.88	94.83	25.64	88.13	37.75	77.51	66.98	89.93	33.31
	50	99.14	3.35	94.90	25.12	88.02	37.64	77.57	66.54	89.91	33.16
	60	99.02	3.96	94.97	24.48	87.91	37.20	77.54	66.26	89.86	32.98
	70	98.90	4.47	95.01	24.18	87.78	37.53	77.47	66.23	89.79	33.10
	80	98.79	5.05	95.02	24.02	87.63	37.64	77.35	65.75	89.70	33.12
	90	98.70	5.50	95.01	24.15	87.46	38.42	77.13	65.58	89.57	33.41
	100	98.60	5.85	94.98	24.11	87.20	38.71	76.75	65.28	89.38	33.49
NegRefine	15	99.57	1.47	95.00	23.85	90.65	33.18	81.92	61.96	91.79	30.12
	20	99.51	1.80	95.08	23.44	90.45	33.26	81.76	61.81	91.70	30.08
	30	99.37	2.28	95.26	22.97	90.56	32.97	81.95	61.11	91.78	29.83
	40	99.21	2.77	95.28	22.86	90.29	32.82	81.84	60.41	91.66	29.71
	50	99.07	3.47	95.31	22.74	90.01	33.41	81.35	60.21	91.44	29.96
	60	98.91	4.05	95.30	22.56	89.74	33.70	81.19	60.39	91.28	30.18
	70	98.78	4.49	95.27	22.64	89.42	34.00	80.57	60.80	91.01	30.48
	80	98.64	4.95	95.20	23.01	89.11	34.40	80.14	60.96	90.77	30.83
	90	98.47	5.52	95.05	23.38	88.71	35.06	79.51	60.99	90.43	31.24
	100	98.27	6.20	94.78	24.38	88.15	36.28	78.55	61.77	89.94	32.16

Table 7. Ablation study on different values of the initial negative label selection percentage p using the ImageNet-1K in-distribution benchmark. For each method, the value of p that achieves the best average FPR95 is highlighted in gray. Across all values of p, our method (NegRefine) consistently outperforms CSP and NegLabel in both average AUROC and FPR95.

LLM	iNaturalist		OpenImage-O		Clean		NINCO		Average	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
Qwen2.5-14B-Instruct	99.57	1.47	**95.00**	**23.85**	**90.65**	**33.18**	81.92	61.96	91.79	30.12
Qwen2.5-7B-Instruct-1M	99.55	1.61	94.98	23.94	90.60	33.63	81.83	61.86	91.74	30.26
Mistral-7B-Instruct-v0.2	99.60	1.36	94.82	24.34	90.63	33.37	**82.92**	**61.14**	**91.99**	30.05
Meta-Llama-3-8B-Instruct	**99.61**	**1.32**	94.92	24.04	90.56	33.48	82.61	61.31	91.93	**30.04**

Table 8. Ablation study on different LLMs for negative label filtering using the ImageNet-1K in-distribution benchmark.

Architecture	Method	iNaturalist		OpenImage-O		Clean		NINCO		Average	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
ResNet50	NegLabel	99.24	2.88	92.00	33.35	84.45	45.52	74.20	73.53	87.47	38.82
	CSP	**99.46**	**1.95**	91.83	33.91	86.08	42.06	75.53	72.86	88.22	37.70
	NegRefine	99.38	2.30	**92.57**	**31.05**	**88.26**	**38.97**	**80.73**	**65.62**	**90.23**	**34.48**
ResNet101	NegLabel	99.27	3.11	90.92	37.41	84.83	46.34	76.80	72.93	87.95	39.95
	CSP	99.47	2.04	91.90	35.58	85.81	43.43	75.92	72.25	88.28	38.33
	NegRefine	**99.59**	**1.49**	**93.75**	**27.52**	**88.43**	**38.23**	**80.17**	**66.06**	**90.48**	**33.33**
ResNet50x4	NegLabel	99.45	2.27	91.93	34.83	86.28	43.24	78.33	69.96	89.00	37.58
	CSP	99.65	1.48	93.38	31.62	87.84	39.01	79.14	68.25	90.00	35.09
	NegRefine	**99.66**	**1.25**	**94.32**	**26.16**	**89.75**	**35.51**	**82.27**	**61.53**	**91.50**	**31.11**
ResNet50x16	NegLabel	99.48	2.00	92.50	33.82	88.12	40.41	80.56	65.57	90.17	35.45
	CSP	**99.68**	**1.25**	93.91	30.47	89.92	38.20	82.20	63.22	91.43	33.28
	NegRefine	99.60	1.42	**94.68**	**25.59**	**91.19**	**34.40**	**84.18**	**57.18**	**92.41**	**29.65**
ResNet50x64	NegLabel	99.63	1.46	93.68	30.58	90.15	37.42	86.98	54.41	92.61	30.97
	CSP	**99.69**	**1.19**	93.95	30.21	91.48	33.92	87.18	52.94	93.08	29.57
	NegRefine	99.60	1.41	**94.52**	**26.45**	**92.14**	**30.90**	**88.26**	**47.33**	**93.63**	**26.52**
ViT-B/32	NegLabel	99.11	3.73	92.87	31.26	85.20	44.16	75.30	71.45	88.12	37.65
	CSP	**99.46**	2.37	93.37	31.01	87.42	40.15	76.96	70.62	89.30	36.04
	NegRefine	99.45	**2.20**	**94.22**	**26.11**	**89.90**	**35.47**	**82.38**	**60.70**	**91.49**	**31.12**
ViT-B/16	NegLabel	99.49	1.91	93.74	28.62	86.79	41.22	77.30	68.70	89.33	35.11
	CSP	**99.60**	1.54	94.09	28.94	88.32	38.75	77.88	68.65	89.97	34.47
	NegRefine	99.57	**1.47**	**95.00**	**23.85**	**90.65**	**33.18**	**81.92**	**61.96**	**91.79**	**30.12**
ViT-L/14	NegLabel	99.53	1.77	94.26	27.55	89.09	38.38	81.98	62.59	91.22	32.57
	CSP	**99.72**	**1.21**	95.02	25.59	91.52	34.22	84.94	57.20	92.80	29.55
	NegRefine	99.67	1.25	**95.96**	**19.92**	**92.74**	**28.43**	**87.62**	**47.40**	**94.00**	**24.25**
ViT-L/14-336px	NegLabel	99.67	1.31	94.26	27.28	89.64	36.87	83.70	61.06	91.82	31.63
	CSP	**99.79**	**0.86**	94.90	26.06	91.99	32.19	86.38	55.90	93.27	28.75
	NegRefine	99.71	1.01	**95.73**	**20.46**	**93.16**	**26.92**	**88.81**	**44.68**	**94.35**	**23.27**

Table 9. Ablation study comparing our method (NegRefine) with CSP and NegLabel across different CLIP architectures using the ImageNet-1K in-distribution benchmark.

S_{MM} Score	α	OOD Datasets: iNaturalist AUROC ↑	iNaturalist FPR95 ↓	OpenImage-O AUROC ↑	OpenImage-O FPR95 ↓	Clean AUROC ↑	Clean FPR95 ↓	NINCO AUROC ↑	NINCO FPR95 ↓	Average AUROC ↑	Average FPR95 ↓
$\max_{i,j}\ (\text{sim}(x,t_{i,j})/\tau - \text{sim}(x,\tilde{y}_j)/\tau)$	1	98.10	8.51	92.57	31.85	89.60	37.79	81.63	62.93	90.47	35.27
	2	97.71	10.35	92.30	33.10	89.39	38.60	81.55	63.34	90.24	36.35
	3	97.54	11.10	92.20	33.64	89.31	39.01	81.53	63.42	90.15	36.79
	4	97.45	11.55	92.14	33.96	89.27	39.30	81.51	63.39	90.09	37.05
$\max_{i,j}\ \frac{\text{sim}(x,t_{i,j})/\tau}{\text{sim}(x,t_{i,j})/\tau+\text{sim}(x,\tilde{y}_j)/\tau}$	1	**99.65**	1.26	94.85	24.16	90.50	34.66	80.76	65.65	91.44	31.43
	2	**99.65**	**1.24**	94.66	23.65	90.68	33.92	81.11	64.78	91.53	30.90
	3	99.63	1.29	94.49	**23.62**	**90.71**	33.70	81.26	64.15	91.52	30.69
	4	99.62	1.32	94.34	23.63	90.69	33.15	81.33	63.54	91.49	30.41
$\max_{i,j}\ \frac{e^{\text{sim}(x,t_{i,j})/\tau}}{e^{\text{sim}(x,t_{i,j})/\tau}+e^{\text{sim}(x,\tilde{y}_j)/\tau}}$	1	99.62	1.38	**95.07**	23.63	90.39	33.63	80.77	63.78	91.46	30.61
	2	99.57	1.47	95.00	23.85	90.65	33.18	**81.92**	**61.96**	**91.79**	**30.12**
	3	99.50	1.67	94.85	24.30	90.65	**33.04**	81.49	62.61	91.62	30.41
	4	99.43	1.76	94.72	24.85	90.66	33.19	81.63	62.11	91.61	30.48
$\text{avg}_{i,j}\ \frac{e^{\text{sim}(x,t_{i,j})/\tau}}{e^{\text{sim}(x,t_{i,j})/\tau}+e^{\text{sim}(x,\tilde{y}_j)/\tau}}$	1	98.95	3.79	91.01	32.12	87.81	39.04	72.67	70.14	87.61	36.27
	2	98.36	6.21	89.85	36.06	86.96	41.44	71.93	71.86	86.77	38.89
	3	97.86	8.48	89.23	38.22	86.48	43.28	71.58	73.07	86.29	40.76
	4	97.46	10.63	88.83	39.75	86.16	44.83	71.38	73.55	85.96	42.19
$\sum_{i,j}\ \frac{e^{\text{sim}(x,t_{i,j})/\tau}}{e^{\text{sim}(x,t_{i,j})/\tau}+e^{\text{sim}(x,\tilde{y}_j)/\tau}}$	1	95.41	20.08	87.31	45.84	84.92	49.65	70.69	76.32	84.58	47.97
	2	95.08	21.73	87.11	46.86	84.75	50.68	70.60	76.73	84.38	49.00
	3	94.97	22.24	87.04	47.16	84.69	51.16	70.57	76.83	84.32	49.35
	4	94.91	22.48	87.00	47.31	84.66	51.31	70.56	76.86	84.28	49.49

Table 10. Ablation study on different design choices for the S_{MM} score function across various α values. The configuration adopted in the main paper is highlighted in gray.

k	OOD Datasets									
	iNaturalist		OpenImage-O		Clean		NINCO		Average	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
1	98.40	7.63	92.03	34.87	88.84	40.66	80.71	64.29	89.99	36.86
2	99.21	2.82	93.94	28.19	90.24	35.43	82.11	**61.31**	91.38	31.94
3	99.43	2.00	94.59	25.58	90.53	**32.93**	82.10	61.87	91.66	30.59
4	99.53	1.71	94.94	24.30	**90.71**	32.97	**82.24**	61.67	**91.86**	30.16
5	99.57	1.47	95.00	23.85	90.65	33.18	81.92	61.96	91.79	**30.12**
6	99.59	1.50	95.12	23.73	90.61	33.30	82.11	62.45	**91.86**	30.25
7	99.61	1.46	95.17	23.54	90.59	33.19	82.01	62.59	91.84	30.20
8	99.61	1.40	95.20	23.22	90.51	33.19	81.89	62.71	91.80	30.13
9	99.63	1.37	95.22	23.22	90.44	33.48	81.81	62.93	91.77	30.25
10	99.63	1.38	**95.23**	23.33	90.42	33.30	81.75	63.24	91.76	30.31
15	**99.65**	1.35	**95.23**	**23.17**	90.22	33.70	81.43	64.10	91.63	30.58
20	**99.65**	**1.32**	95.21	23.34	90.03	33.78	81.20	64.09	91.52	30.63

Table 11. Ablation study on different values of k (the number of top in-distribution and negative labels used to create multi-matching texts). In the paper, we adopted $k = 5$ (highlighted in gray).

In-distribution data	Method	OOD Datasets									
		iNaturalist		SUN		Places		Textures		Average	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
Stanford-Cars	NegLabel	99.99	0.01	99.99	0.01	**99.99**	0.03	99.99	0.01	99.99	0.02
	CSP	**100.0**	**0.00**	**100.0**	**0.00**	**99.99**	**0.02**	**100.0**	**0.00**	**100.0**	**0.01**
	NegRefine	**100.0**	**0.00**	**100.0**	0.01	**99.99**	0.07	**100.0**	**0.00**	**100.0**	0.02
CUB-200	NegLabel	**99.96**	0.18	**99.99**	**0.02**	**99.90**	**0.33**	99.99	0.01	**99.96**	**0.14**
	CSP	**99.96**	**0.16**	**99.99**	0.03	99.88	0.37	**100.0**	**0.00**	**99.96**	**0.14**
	NegRefine	99.92	0.19	99.95	0.14	99.68	1.04	99.99	0.02	99.89	0.35
Oxford-Pet	NegLabel	99.99	0.01	99.99	0.02	**99.96**	**0.17**	**99.97**	**0.11**	**99.98**	**0.08**
	CSP	**100.0**	**0.00**	**100.0**	**0.00**	**99.96**	0.21	**99.97**	0.14	**99.98**	0.09
	NegRefine	**100.0**	**0.00**	99.99	0.01	99.95	0.24	99.96	0.12	99.97	0.09
Food-101	NegLabel	99.99	0.01	99.99	0.01	**99.99**	0.01	99.60	1.61	99.89	0.41
	CSP	**100.0**	**0.00**	**100.0**	**0.00**	**99.99**	0.01	99.63	1.40	**99.91**	0.35
	NegRefine	99.99	**0.00**	99.98	**0.00**	99.98	**0.00**	**99.68**	**1.37**	**99.91**	**0.34**

Table 12. OOD detection performance using other in-distribution datasets.