

BENCHMARKING OPEN-SET RECOGNITION BEYOND VISION-LANGUAGE PRE-TRAINING

Anonymous authors

Paper under double-blind review

This appendix is organized as follows:

- In Section A, we include dataset split details, including the FGVC Aircraft dataset splits in Section A.1, iNaturalist dataset configurations in Section A.2, and Plant with Disease dataset settings in Section A.3.
- In Section B, we discuss additional implementation details, including baseline implementation specifications for SD-LRT and CLIP-based models.
- In Section C, we discuss comprehensive experimental results, including FGVC Aircraft results with classical settings in Section C.1, detailed FGVC Aircraft performance across all split configurations in Section C.2, complete iNaturalist results in Section C.3, and VisA dataset category-wise performance in Section C.4.
- In Section D, we show method extensions and applications, including the adaptation of SD-LRT for anomaly detection tasks with modified pipeline and error map processing approaches.
- In Section E, we present baseline analysis and zero-shot performance evaluation of vision-language models across multiple canonical datasets, demonstrating the generalization capabilities and limitations of CLIP, OpenCLIP, and diffusion-based classifiers.
- In Section F and G, we discuss related work covering the foundations of open-set recognition and the current limitations of our work and broader impacts, respectively.

A DATASET SPLIT DETAILS

In our experiments, we evaluate several methods on multiple datasets including FGVC Aircraft, iNaturalist, and Plant with Disease. Here we detail the split configuration for open-set and closed-set classes.

A.1 FGVC AIRCRAFT DATASET

We created four different split settings for evaluating methods. The open-set classes of different settings are listed below:

Intra-12. Open-Set classes: ‘707-320’, ‘747-100’, ‘A310’, ‘A340-600’, ‘BAE 146-200’, ‘Cessna 172’, ‘DC-10’, ‘F-16A/B’, ‘Falcon 2000’, ‘DHC-1’, ‘E-170’, ‘MD-11’.

Inter-12. Open-Set classes: ‘737-200’, ‘737-300’, ‘737-400’, ‘737-500’, ‘737-600’, ‘737-700’, ‘737-800’, ‘737-900’, ‘747-100’, ‘747-200’, ‘747-300’, ‘747-400’.

Inter-32. Open-Set classes: ‘737-200’, ‘737-300’, ‘737-400’, ‘737-500’, ‘737-600’, ‘737-700’, ‘737-800’, ‘737-900’, ‘747-100’, ‘747-200’, ‘747-300’, ‘747-400’, ‘757-200’, ‘757-300’, ‘767-200’, ‘767-300’, ‘767-400’, ‘777-200’, ‘777-300’, ‘A300B4’, ‘A310’, ‘A318’, ‘A319’, ‘A320’, ‘A321’, ‘A330-200’, ‘A330-300’, ‘A340-200’, ‘A340-300’, ‘A340-500’, ‘A340-600’, ‘A380’.

Intra-32. Open-Set classes: ‘707-320’, ‘A310’, ‘A340-600’, ‘BAE 146-200’, ‘Cessna 172’, ‘DC-10’, ‘F-16A/B’, ‘Falcon 2000’, ‘DHC-1’, ‘E-170’, ‘MD-11’, ‘727-200’, ‘737-200’, ‘Yak-42’, ‘Tu-154’, ‘Spitfire’, ‘Saab 2000’, ‘L-1011’, ‘Gulfstream V’, ‘Global Express’, ‘ERJ 145’, ‘DH-82’, ‘CRJ-900’, ‘ATR-72’, ‘An-12’, ‘Challenger 600’, ‘747-200’, ‘777-200’, ‘Il-76’, ‘Metroliner’, ‘Hawk T1’, ‘Beechcraft 1900’.

A.2 INATURALIST

We created five different split settings for evaluating methods. The open- and close-set classes of different settings are listed below:

Intra-Order. Open-Set classes: ‘Cheiracanthium’, ‘Nyssus’, ‘Dysdera’. Close-Set classes: ‘Cytorphora’, ‘Badumna’, ‘Neriene’, ‘Oxyopes’, ‘Holocnemus’, ‘Dolomedes’, ‘Sidymella’.

Same-Family. Open-Set classes: ‘Nephila’, ‘Paraphidippus’, ‘Neoscona’. Close-Set classes: ‘Araneus’, ‘Cyclosa’, ‘Eriophora’, ‘Attulus’, ‘Hasarius’, ‘Metacyrba’.

Cross-Family. Open-Set classes: ‘Herpyllus’, ‘Rabidosa’, ‘Misumena’. Close-Set classes: ‘Araneus’, ‘Cyclosa’, ‘Eriophora’, ‘Attulus’, ‘Hasarius’, ‘Metacyrba’.

Intra-Family. Open-Set classes: ‘Araneus’, ‘Cyclosa’, ‘Eriophora’. Close-Set classes: ‘Aculepeira’, ‘Argiope’, ‘Austracantha’, ‘Cyrtophora’, ‘Gasteracantha’, ‘Larinioides’, ‘Mangora’.

Inter-Family. Open-Set classes: ‘Anasaitis’, ‘Attulus’, ‘Hasarius’. Close-Set classes: ‘Aculepeira’, ‘Argiope’, ‘Austracantha’, ‘Cyrtophora’, ‘Gasteracantha’, ‘Larinioides’, ‘Mangora’.

A.3 PLANT WITH DISEASE

We created two different split settings for evaluating methods. The open-set classes of different settings are listed below:

Intra-Plant. Open-Set classes: ‘Tomato Yellow Leaf Curl Virus’, ‘Grape Leaf blight Isariopsis Leaf Spot’, ‘Strawberry Leaf scorch’, ‘Cherry Powdery mildew’, ‘Corn Cercospora leaf spot Gray leaf spot’, ‘Apple Cedar apple rust’.

Inter-Plant. Open-Set classes: ‘Tomato Bacterial spot’, ‘Tomato Early blight’, ‘Tomato Late blight’, ‘Tomato Leaf Mold’, ‘Tomato Septoria leaf spot’, ‘Tomato Spider mites Two-spotted spider mite’, ‘Tomato Target Spot’, ‘Tomato Yellow Leaf Curl Virus’, ‘Tomato mosaic virus’, ‘Tomato healthy’.

B IMPLEMENTATION DETAILS OF BASELINES

SD-LRT. We used SD 2.0 following (Li et al., 2023b; Yue et al., 2024). For LoRA matrices rank, we used 16 for each dataset. For a fair comparison, We use ”a photo of [c], a type of [SC]” for CLIP-based baseline methods on FGVC Aircraft, where [c] denotes the name of class c and [SC] denotes the dataset-specific super-class name. We used the same prompt as the original papers for the anomaly detection methods. For SD-LRT, we used ”a photo of [c], a type of [SC]” for both SD models with and without LoRA.

CLIP-based models. We consider two variants of CLIP models that are publicly available: CLIP (Radford et al., 2021) with ViT-B/16 backbone trained on 400M image-caption pairs, and OpenCLIP (Ilharco et al., 2021) with ViT-H/14 backbone trained on LAION-2B (Schuhmann et al., 2022). In zero-shot mode, these models classify images by comparing their visual features with predefined text prompts (e.g., “a photo of an aircraft”) and selecting the class with highest image-prompt similarity.

C DETAILED EXPERIMENT RESULTS

C.1 FGVC AIRCRAFT RESULT WITH CLASSICAL SETTING

To ensure fair comparison, we evaluate SD-LRT on the FGVC Aircraft dataset under both Medium and Hard difficulty levels following classical open-set recognition setting (Vaze et al., 2021). As shown in Table A1, VLMs demonstrate similar performance patterns across both difficulty settings, suggesting that traditional difficulty metrics may **not** effectively capture the challenges faced by these models.

As observed in our main findings, CLIP-based methods exhibit **unstable** open-set performance as shot numbers increase. In contrast, SD-LRT demonstrates consistent improvement in both closed-

Methods	Medium					Hard				
	1	2	4	8	16	1	2	4	8	16
CoCoOp	32.1 / 45.1	39.9 / 44.2	42.6 / 44.4	42.2 / 41.4	45.6 / 42.5	32.1 / 63.6	39.9 / 59.1	42.6 / 60.7	42.2 / 72.1	45.6 / 67.0
MaPLe	39.2 / 40.5	39.7 / 54.4	44.2 / 47.4	46.7 / 44.2	54.2 / 48.8	39.2 / 61.8	39.7 / 59.9	44.2 / 63.0	46.7 / 44.2	54.2 / 61.8
LoCoOp	28.3 / 42.8	42.1 / 42.5	39.4 / 46.1	36.8 / 42.4	48.7 / 48.3	28.2 / 65.5	35.5 / 62.3	45.2 / 64.2	47.0 / 64.6	49.7 / 73.1
Tip-Adapter	53.2 / 62.5	59.1 / 64.4	59.9 / 60.9	61.6 / 60.8	63.9 / 58.5	53.2 / 58.1	59.1 / 59.3	59.9 / 62.5	63.9 / 63.5	63.9 / 61.5
Tip-Adapter-F	57.5 / 62.4	61.5 / 62.7	62.2 / 63.6	67.2 / 60.8	71.8 / 64.4	57.5 / 57.9	61.5 / 60.8	62.2 / 64.2	67.2 / 61.6	71.8 / 63.3
SD-LRT	56.4 / 49.5	61.9 / 49.5	72.9 / 57.7	75.0 / 72.9	79.0 / 79.3	56.4 / 58.6	61.9 / 60.8	72.9 / 67.2	75.0 / 75.4	79.0 / 78.7

Table A1: Performance comparison (F1-score/AUROC %) under the classic medium/hard difficulty settings for open-set recognition on FGVC Aircraft dataset, across varying few-shot scenarios. SD-LRT consistently outperforms existing methods, particularly in 4/8/16-shot settings where it achieves superior results in both metrics (e.g., F1: 72.9/75.0/79.0, AUROC: 67.2/75.4/78.7 vs next best F1: 62.2/67.2/71.8, AUROC: 64.2/72.1/67.0 in medium setting).

set and open-set performance across increasing shots under both difficulty settings. This **stable** improvement is particularly evident in 4-, 8-, and 16-shot scenarios, achieving comparable performance (79.0/79.3 and 79.0/78.7 for Medium and Hard at 16-shot) across different difficulty levels. The consistent performance improvement and minimal gap between difficulty settings validate SD-LRT’s robustness in both closed-set and open-set recognition tasks.

C.2 DETAILED RESULT ON FGVC AIRCRAFT

Methods	Intra-12					Inter-12				
	1	2	4	8	16	1	2	4	8	16
CocoOp	18.8 / 58.7	21.8 / 59.2	24.5 / 62.1	27.0 / 59.3	27.0 / 62.4	24.5 / 67.6	25.6 / 76.9	25.8 / 75.4	29.7 / 76.4	30.6 / 74.4
MaPLe	25.7 / 56.9	21.6 / 60.1	23.2 / 66.4	27.8 / 58.6	27.6 / 59.4	21.1 / 71.2	24.5 / 76.7	26.5 / 75.9	28.9 / 73.1	31.6 / 74.1
LoCoOp	24.9 / 55.4	27.8 / 56.7	32.7 / 54.8	35.9 / 59.5	41.1 / 61.1	24.7 / 72.5	29.5 / 71.0	36.0 / 74.7	40.3 / 71.5	44.6 / 76.3
Tip-Adapter (Clip)	19.6 / 49.2	21.1 / 53.1	23.9 / 51.9	23.6 / 53.1	26.7 / 54.3	19.5 / 65.0	23.2 / 65.6	25.5 / 66.5	28.1 / 62.6	30.1 / 68.3
Tip-Adapter-F (Clip)	20.6 / 46.7	24.2 / 49.4	27.3 / 49.4	30.4 / 52.4	34.5 / 53.5	22.9 / 61.7	26.0 / 58.6	31.3 / 62.5	34.4 / 61.2	37.9 / 64.2
Tip-Adapter (B/32)	23.1 / 52.3	27.1 / 52.5	30.6 / 54.2	31.7 / 56.7	37.2 / 58.3	23.5 / 62.9	28.9 / 67.1	31.8 / 66.8	37.7 / 69.8	39.2 / 69.6
Tip-Adapter-F (B/32)	27.0 / 52.1	29.3 / 47.5	33.3 / 49.7	37.4 / 49.0	46.1 / 50.9	27.3 / 65.8	31.2 / 65.7	34.5 / 61.0	40.6 / 68.7	48.2 / 69.2
Tip-Adapter (H/14)	43.7 / 54.3	46.8 / 57.3	49.6 / 56.0	52.1 / 56.0	56.6 / 57.6	46.9 / 78.8	49.1 / 80.7	51.7 / 81.4	55.9 / 82.6	59.9 / 81.1
Tip-Adapter-F (H/14)	45.7 / 53.3	50.3 / 50.9	53.4 / 44.1	57.4 / 48.5	64.0 / 49.1	48.4 / 82.1	52.6 / 84.7	56.0 / 82.4	62.4 / 82.2	67.4 / 81.8
Ours	45.3 / 45.9	53.5 / 47.0	65.3 / 59.3	70.9 / 70.9	78.5 / 76.5	52.3 / 47.5	54.8 / 54.4	67.5 / 59.1	74.8 / 82.6	81.2 / 87.8
Methods	Intra-32					Inter-32				
	1	2	4	8	16	1	2	4	8	16
CocoOp	19.8 / 52.2	20.5 / 51.3	23.1 / 53.7	25.2 / 52.3	27.0 / 52.1	28.6 / 82.2	31.6 / 86.7	31.8 / 91.0	36.0 / 88.1	39.3 / 89.7
MaPLe	21.5 / 52.8	22.3 / 55.6	23.4 / 49.8	23.1 / 54.8	26.7 / 53.6	29.7 / 90.6	31.4 / 89.0	32.2 / 87.4	37.4 / 87.1	38.5 / 90.7
LoCoOp	24.8 / 51.7	26.8 / 50.6	37.8 / 51.7	38.9 / 57.6	43.5 / 53.1	27.3 / 82.6	23.4 / 76.6	41.7 / 82.3	44.1 / 85.4	50.5 / 86.2
Tip-Adapter (CLIP)	17.6 / 46.0	18.9 / 45.0	21.3 / 50.4	24.6 / 54.1	25.2 / 53.9	22.2 / 59.7	24.0 / 62.7	28.5 / 67.9	32.3 / 68.0	32.8 / 71.3
Tip-Adapter-F (CLIP)	17.9 / 48.2	20.8 / 48.5	23.5 / 44.8	28.9 / 48.2	32.7 / 51.8	24.6 / 48.5	28.5 / 51.4	32.1 / 47.1	38.3 / 57.3	39.4 / 65.3
Tip-Adapter (B/32)	21.8 / 51.8	26.2 / 50.4	28.4 / 53.5	32.1 / 53.7	35.9 / 53.7	28.2 / 63.9	31.1 / 65.5	35.5 / 66.2	39.7 / 70.8	43.5 / 73.0
Tip-Adapter-F (B/32)	25.2 / 49.8	27.5 / 47.7	29.9 / 46.5	36.1 / 46.6	41.1 / 51.1	30.5 / 63.1	31.8 / 56.4	34.2 / 56.2	40.6 / 64.4	49.7 / 68.0
Tip-Adapter (H/14)	45.2 / 58.7	48.1 / 60.1	50.7 / 61.3	54.8 / 64.3	57.6 / 64.3	53.6 / 66.6	55.0 / 70.2	62.2 / 72.1	63.4 / 79.1	66.3 / 77.6
Tip-Adapter-F (H/14)	48.4 / 56.9	50.6 / 54.8	53.9 / 54.2	57.8 / 58.8	62.1 / 60.6	54.9 / 63.9	57.9 / 64.9	62.4 / 66.9	67.7 / 75.3	71.6 / 72.7
SD-LRT	47.6 / 46.4	55.2 / 48.7	66.7 / 59.1	71.2 / 71.4	76.7 / 77.2	56.4 / 51.0	57.9 / 59.2	71.2 / 62.1	80.7 / 88.3	85.1 / 91.3

Table A2: Performance comparison on FGVC Aircraft dataset under four settings combining different open-set ratios (12 or 32 open-set classes against 88/68 closed-set classes) and semantic relationships (Intra/Inter-family).

In Table A2, we provide the full results of our experiments on FGVC Aircraft dataset. On FGVC-Aircraft, by varying backbones for both Tip-Adapter and Tip-Adapter-F, we observe that increased backbone parameters lead to simultaneous improvements in closed-set classification and open-set recognition performance. Across four difficulty settings, stronger backbones consistently outperform weaker ones, with ViT-H/14 achieving state-of-the-art results while ViT-B/32 backbone showing inferior performance. Backbone choice does not affect other experimental patterns. With identical backbone, Tip-Adapter-F consistently excels in closed-set classification but underperforms Tip-Adapter in open-set recognition. As stronger backbones are pre-trained on larger datasets, the enhanced pre-training data likely enables better approximation of semantic boundaries between classes.

Intra-Order					
Methods	1	2	4	8	16
CoCoOp	19.3 / 62.2	37.1 / 47.9	45.4 / 52.6	58.4 / 58.5	70.9 / 64.7
MaPLe	27.6 / 52.2	47.2 / 53.7	69.6 / 63.4	72.5 / 68.8	84.5 / 80.6
LoCoOp	30.9 / 63.7	30.1 / 64.2	40.9 / 69.2	56.8 / 65.8	71.7 / 70.1
Tip-Adapter	50.5 / 55.4	57.7 / 72.8	71.2 / 67.3	75.2 / 76.3	81.9 / 80.6
Tip-Adapter-F	53.2 / 58.2	68.8 / 57.0	72.1 / 64.7	75.9 / 63.1	84.8 / 74.3
Ours	40.1 / 68.8	64.7 / 74.8	70.8 / 77.9	73.4 / 79.0	82.6 / 81.5
Same-Family					
Methods	1	2	4	8	16
CoCoOp	14.4 / 49.5	19.9 / 48.0	30.9 / 57.8	44.1 / 60.0	53.2 / 68.4
MaPLe	31.1 / 48.9	24.6 / 63.2	47.5 / 62.7	58.4 / 62.7	72.7 / 65.9
LoCoOp	23.1 / 54.5	32.1 / 52.7	38.4 / 56.7	40.5 / 57.6	60.8 / 61.1
Tip-Adapter	27.5 / 66.7	39.8 / 51.1	45.8 / 63.1	48.3 / 57.8	74.0 / 64.9
Tip-Adapter-F	31.1 / 60.4	43.3 / 66.8	46.3 / 63.0	59.6 / 66.3	71.1 / 67.9
Ours	26.0 / 57.1	36.0 / 58.3	51.1 / 62.2	62.6 / 64.1	76.2 / 68.7
Cross-Family					
Methods	1	2	4	8	16
CoCoOp	14.4 / 45.9	19.9 / 52.7	30.9 / 56.2	44.1 / 55.8	53.2 / 54.2
MaPLe	31.1 / 46.9	24.6 / 68.7	47.5 / 55.4	58.4 / 60.6	72.7 / 58.3
LoCoOp	31.1 / 53.5	32.1 / 49.6	38.4 / 51.5	40.5 / 56.1	60.8 / 59.2
Tip-Adapter	27.5 / 66.7	39.8 / 73.2	45.8 / 63.1	48.3 / 57.9	74.0 / 62.8
Tip-Adapter-F	31.1 / 62.8	43.3 / 60.4	46.3 / 60.9	59.6 / 62.4	71.1 / 62.8
Ours	26.0 / 59.5	36.0 / 62.7	51.1 / 63.1	62.6 / 64.4	76.2 / 64.5
Intra-Family					
Methods	1	2	4	8	16
CoCoOp	24.1 / 57.2	19.6 / 56.0	36.2 / 57.4	40.2 / 58.1	53.6 / 52.7
MaPLe	19.3 / 43.0	21.4 / 52.7	42.0 / 47.3	48.9 / 48.1	50.5 / 54.4
LoCoOp	29.7 / 51.4	22.4 / 54.2	37.4 / 51.6	42.1 / 53.7	47.5 / 51.5
Tip-Adapter	24.3 / 45.7	32.3 / 58.9	51.4 / 58.4	59.8 / 60.0	66.7 / 59.6
Tip-Adapter-F	32.3 / 58.9	41.8 / 51.7	47.7 / 61.0	53.3 / 58.1	62.3 / 55.8
Ours	32.9 / 51.2	37.5 / 57.6	51.5 / 59.3	63.4 / 63.7	75.1 / 65.5
Inter-Family					
Methods	1	2	4	8	16
CoCoOp	24.1 / 51.9	19.6 / 39.9	36.2 / 44.2	40.2 / 37.3	53.6 / 51.9
MaPLe	19.3 / 61.9	21.4 / 63.5	42.0 / 65.8	48.9 / 37.2	50.5 / 59.0
LoCoOp	29.7 / 60.3	22.4 / 62.7	37.4 / 63.6	42.1 / 63.4	47.5 / 66.9
Tip-Adapter	24.3 / 42.3	32.3 / 61.3	51.4 / 60.0	59.8 / 52.3	66.7 / 57.2
Tip-Adapter-F	32.3 / 47.5	41.8 / 52.5	47.7 / 52.5	53.3 / 55.2	62.3 / 59.6
SD-LRT	32.9 / 58.1	37.5 / 60.4	51.5 / 69.9	63.4 / 76.1	75.1 / 79.2

Table A3: Performance comparison on iNaturalist dataset across five semantic settings. Each cell reports closed-set F1 score (%) / open-set AUROC (%) under N-shot settings. Results compare CoCoOp, MaPLe, LoCoOp, Tip-Adapter, Tip-Adapter-F, and SD-LRT.

C.3 DETAILED RESULT ON INATURALIST

In Table A3, we provide the full results of our experiments on iNaturalist dataset.

C.4 DETAILED RESULT ON VISA

In Table. A4, we show detailed results on Visa divided by categories. SD-LRT achieves consistent performance across different categories in Visa dataset, maintaining balanced $F_1 - max$ and AUROC scores across various objects (ranging from food items like cashew and macaroni to industrial components like PCB). This robustness is particularly evident in the 4-shot setting, where we

Category	1		2		4	
	F1-max	AUROC	F1-max	AUROC	F1-max	AUROC
candle	87.9	92.8	90.5	93.4	91.4	94.9
capsules	89.9	93.2	90.5	93.4	93.4	95.2
cashew	79.1	83.4	86.0	89.6	82.6	85.5
chewinggum	89.9	91.6	90.0	91.4	93.4	93.5
fryum	90.4	90.3	89.0	91.0	92.4	92.3
macaroni1	86.7	93.3	90.5	93.4	93.9	95.3
macaroni2	89.9	90.3	86.4	91.3	90.3	92.3
pcb1	89.9	92.6	90.0	92.6	93.4	94.6
pcb2	89.9	92.7	90.0	92.9	93.4	94.7
pcb3	86.9	92.5	90.0	92.5	93.4	94.5
pcb4	88.4	92.7	90.0	93.3	90.4	94.8
pipe_fryum	88.2	90.3	89.0	90.6	91.9	92.3
average	88.1	91.3	89.3	92.1	91.7	93.3

Table A4: Performance on VisA dataset across different object categories. SD-LRT achieves good F1-max and AUROC scores across various object types, from industrial components (e.g., pcb series) to food items (e.g., cashew, macaroni), demonstrating robust anomaly detection capability under different shot settings (1-, 2-, and 4-shot). The balanced performance across categories, with average F1-max/AUROC scores of 91.7/93.3 in 4-shot setting, validates SD-LRT effectiveness in handling diverse anomaly detection scenarios.

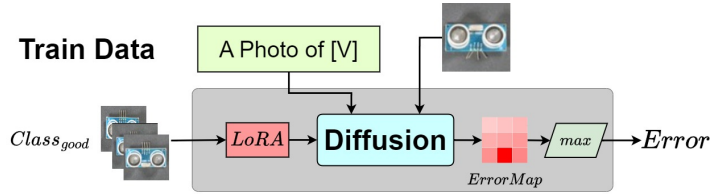


Figure A1: Pipeline of SD-LRT for anomaly detection. We train LoRA parameters for good sample classes while using the maximum error in the error map for anomaly classification.

achieve an average F1-max of 91.7 and AUROC of 93.3, with minimal performance variation across categories. The balanced performance across categories, with average F1-max/AUROC scores of 91.7/93.3 in 4-shot setting, validates SD-LRT’s effectiveness in handling diverse anomaly detection scenarios.

D SD-LRT FOR ANOMALY CLASSIFICATION

For anomaly detection tasks, we employ a similar diffusion classifier as in open-set recognition. However, as shown in Fig. A1, we select the maximum value from the error map rather than the mean as the classification score, since anomaly detection typically involves samples with high overall similarity but significant local differences in anomalous regions. This modification enables SD-LRT to capture localized deviations rather than global differences.

E ZERO-SHOT PERFORMANCE OF VLMS

We evaluate the zero-shot classification performance of CLIP Radford et al. (2021), OpenCLIP Ilharco et al. (2021), and diffusion-based classifier Li et al. (2023a) across several canonical datasets, including Caltech101, Oxford Flowers, and FGVC Aircraft. As shown in Table A5, on many of these benchmarks, VLMS already achieve surprisingly strong performance without any task-specific training, highlighting their powerful generalization capabilities.

However, on the datasets selected in our open-set recognition study—particularly those with fine-grained or domain-specific labels—we observe significantly lower zero-shot performance. This

degradation suggests that VLMs may not have fully captured the fine-grained correspondence between visual features and textual descriptions during pretraining.

Model	Caltech-101	Oxford Pet	Stanford Cars	Oxford Flowers	Food-101	FGVCAircraft	SUN397	DTD	EuroSAT	Plant	iNaturalist
SD Classifier	95.1	87.3	92.3	66.3	77.7	26.4	66.9	38.6	59.4	36.1	21.3
OpenCLIP	94.2	72.8	94.3	69.4	92.7	42.6	75.2	52.5	67.8	32.1	21.1
CLIP	87.5	48.3	87.0	44.3	84.0	19.5	62.4	31.5	44.3	21.5	15.7

Table A5: Performance (%) of CLIP, OpenCLIP and SD classifier across various datasets.

F RELATED WORK

Foundations of Open-set Recognition. Open-set recognition (OSR) is the task of simultaneously classifying samples from known categories while detecting samples from unknown classes not seen during training. The seminal work by Scheirer et al. (2012) formalized this challenge by introducing the concept of open space risk. More recent surveys such as Sun & Dong (2023); Barcina-Blanco et al. (2023) provide a comprehensive taxonomy of OSR methods, spanning threshold-based classifiers, representation learning, and generative models. They also emphasize the practical relevance of OSR in safety-critical domains such as autonomous driving and medical diagnosis.

Discriminative and Generative Approaches Early OSR methods extended closed-set classifiers by introducing specialized mechanisms to identify unknown inputs. OpenMax Bendale & Boulton (2016) pioneered this direction by recalibrating the softmax layer through statistical Extreme Value Theory (EVT), effectively modeling activation patterns of known classes to detect novel categories. Following this trajectory, CROSR Yoshihashi et al. (2019) combined supervised classification with unsupervised reconstruction to create more robust representations that better distinguish between known and unknown classes. Discriminative approaches continued to evolve with methods like C2AE Oza & Patel (2019), which utilizes class-conditional autoencoders to learn compact latent representations for known classes while rejecting unknown samples. Similarly, ARPL Chen et al. (2021) introduced adversarial reciprocal point learning to push decision boundaries of known classes away from potential unknown regions, creating more discriminative feature spaces. Generative approaches have demonstrated substantial promise in addressing OSR challenges. OpenGAN Kong & Ramanan (2021) leverages generative adversarial networks to synthesize unknown class samples, thereby enhancing model robustness to novel inputs. Building on generative modeling, CGDL Sun et al. (2020) employs conditional Gaussian distribution learning to model the feature space of known classes more precisely. Additionally, M2IOSR Sun et al. (2019) utilizes mutual information maximization to develop feature representations that better separate known from unknown classes. More recent innovations include PMAL Lu et al. (2022), which utilizes prototype-based metric adaptation learning to construct more effective decision boundaries for open-set scenarios, and Energy-based models Wang et al. (2024) that assign higher energy values to out-of-distribution samples. These advances collectively demonstrate the community’s progress in developing increasingly effective solutions for open-set recognition tasks.

Vision-Language Models and OSR. The rise of vision-language foundation models (VLMs) such as CLIP and Stable Diffusion has opened new possibilities for OSR. Despite their open-vocabulary nature, recent studies Miller et al. (2024) show that VLMs tend to assign high confidence to out-of-distribution inputs, necessitating new strategies for unknown detection. Prompt-tuning Zhou et al. (2022), knowledge distillation Kim et al. (2024), and contrastive fine-tuning have been explored to improve robustness. Additionally, hierarchical and multi-label OSR methods Wang et al. (2024); Hannum et al. (2024) extend the OSR paradigm to more complex real-world scenarios, reflecting a growing need for structured and generalizable open-world learning systems.

G LIMITATIONS

While our study provides a comprehensive evaluation of open-set recognition (OSR) models across several benchmarks, we acknowledge three key limitations.

First, our evaluation primarily focuses on two canonical metrics: F1 score and AUROC. Although these are widely used in prior OSR literature, additional metrics such as Area Under the Precision-Recall Curve (AUPR), Open Set Classification Rate (OSCR) Dhamija et al. (2018), and False Positive Rate at 95% True Positive Rate (FPR@95) Hendrycks & Gimpel (2016) offer complementary insights. In future work, we plan to include a more complete evaluation using a broader set of OSR metrics.

Second, the SD-LRT model, which is built upon Stable Diffusion and leverages LoRA modules for class-conditional adaptation, requires more computational resources than CLIP-based methods. This becomes particularly significant when scaling to a large number of categories, where the number of LoRA adapters increases linearly with the class set.

Third, although we focus on open-set recognition, we note that out-of-distribution (OOD) detection is a closely related problem. Some OOD detection techniques, such as energy-based models Liu et al. (2020) or Mahalanobis distance Lee et al. (2018), may also be effective for OSR. We leave the integration and unification of OOD detection and OSR within a common framework as a promising direction for future work.

H USING OF LLMs

In the preparation of this manuscript, we utilized Large Language Models (LLMs) solely for language polishing and editing purposes. Specifically, LLMs were employed to:

1. Improve grammatical accuracy and sentence structure
2. Enhance clarity and readability of the text
3. Refine word choice and expression
4. Ensure consistency in academic writing style

REFERENCES

- Marcos Barcina-Blanco, Jesus L Lobo, Pablo Garcia-Bringas, and Javier Del Ser. Managing the unknown: a survey on open set recognition and tangential areas. *arXiv preprint arXiv:2312.08785*, 2023.
- Abhijit Bendale and Terrance E Boulton. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1563–1572, 2016.
- Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081, 2021.
- Akshay Raj Dhamija, Manuel Günther, and Terrance Boulton. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018.
- Andrew Hannum, Max Conway, Mario Lopez, and André Harrison. Data-driven hierarchical open set recognition. *arXiv preprint arXiv:2411.02635*, 2024.
- Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.
- Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL <https://doi.org/10.5281/zenodo.5143773>.
- Seongyeop Kim, Hyung-Il Kim, and Yong Man Ro. Improving open set recognition via visual prompts distilled from common-sense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2786–2794, 2024.
- Shu Kong and Deva Ramanan. Opendata: Open-set recognition via open data generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 813–822, 2021.

- Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.
- Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier, 2023a. URL <https://arxiv.org/abs/2303.16203>.
- Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2206–2217, 2023b.
- Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.
- Jing Lu, Yunlu Xu, Hao Li, Zhazhan Cheng, and Yi Niu. Pmal: Open set recognition via robust prototype mining. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1872–1880, 2022.
- Dimity Miller, Niko Sünderhauf, Alex Kenna, and Keita Mason. Open-set recognition in the age of vision-language models. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.
- Poojan Oza and Vishal M Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2307–2316, 2019.
- 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*, pp. 8748–8763. PMLR, 2021.
- Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7): 1757–1772, 2012.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL <https://openreview.net/forum?id=M3Y74vmsMcY>.
- Jiayin Sun and Qiulei Dong. A survey on open-set image recognition. *arXiv preprint arXiv:2312.15571*, 2023.
- Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 403–412, 2019.
- Xin Sun, Zhenning Yang, Chi Zhang, Keck-Voon Ling, and Guohao Peng. Conditional gaussian distribution learning for open set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13480–13489, 2020.
- Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? 2021.
- Yezhen Wang, Bo Li, Tong Che, Kaiyang Zhou, Ziwei Liu, and Dongsheng Li. Energy-based open-world uncertainty modeling for confidence calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9302–9311.
- Yibo Wang, Jun-Yi Hang, and Min-Ling Zhang. Multi-label open set recognition. *Advances in Neural Information Processing Systems*, 37:5739–5756, 2024.

- Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4016–4025, 2019.
- Zhongqi Yue, Pan Zhou, Richang Hong, Hanwang Zhang, and Qianru Sun. Few-shot learner parameterization by diffusion time-steps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23263–23272, 2024.
- Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022.