NEURAL INFORMATION PROCESSING SYSTEMS

THE UNIVERSITY of ADELAIDE

NAVAL GROUP

Mysteries of the Deep: Role of Intermediate Representations in Out of Distribution Detection 🤿

Ignacio Meza De la Jara, Cristian Rodriguez-Opazo, Damien Teney, Damith Ranasinghe and Ehsan Abbasnejad

THE UNIVERSITY of ADELAIDE

Background

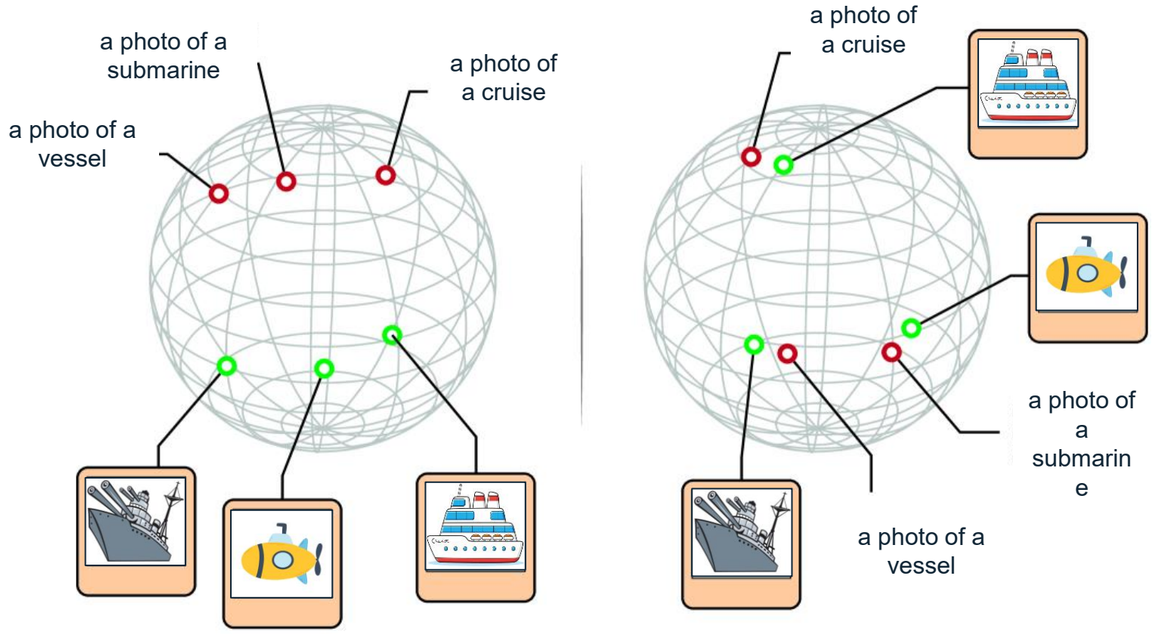
An Overview of OOD Detection

Out-of-Distribution (OOD)

Inference Process with CLIP

- The emergence of **CLIP** enables **zero-shot classification**, where the **model predicts labels by measuring the similarity between image and text embeddings**.

- During inference, **predictions are represented as one-hot encoded vectors based on the classification prompts.**

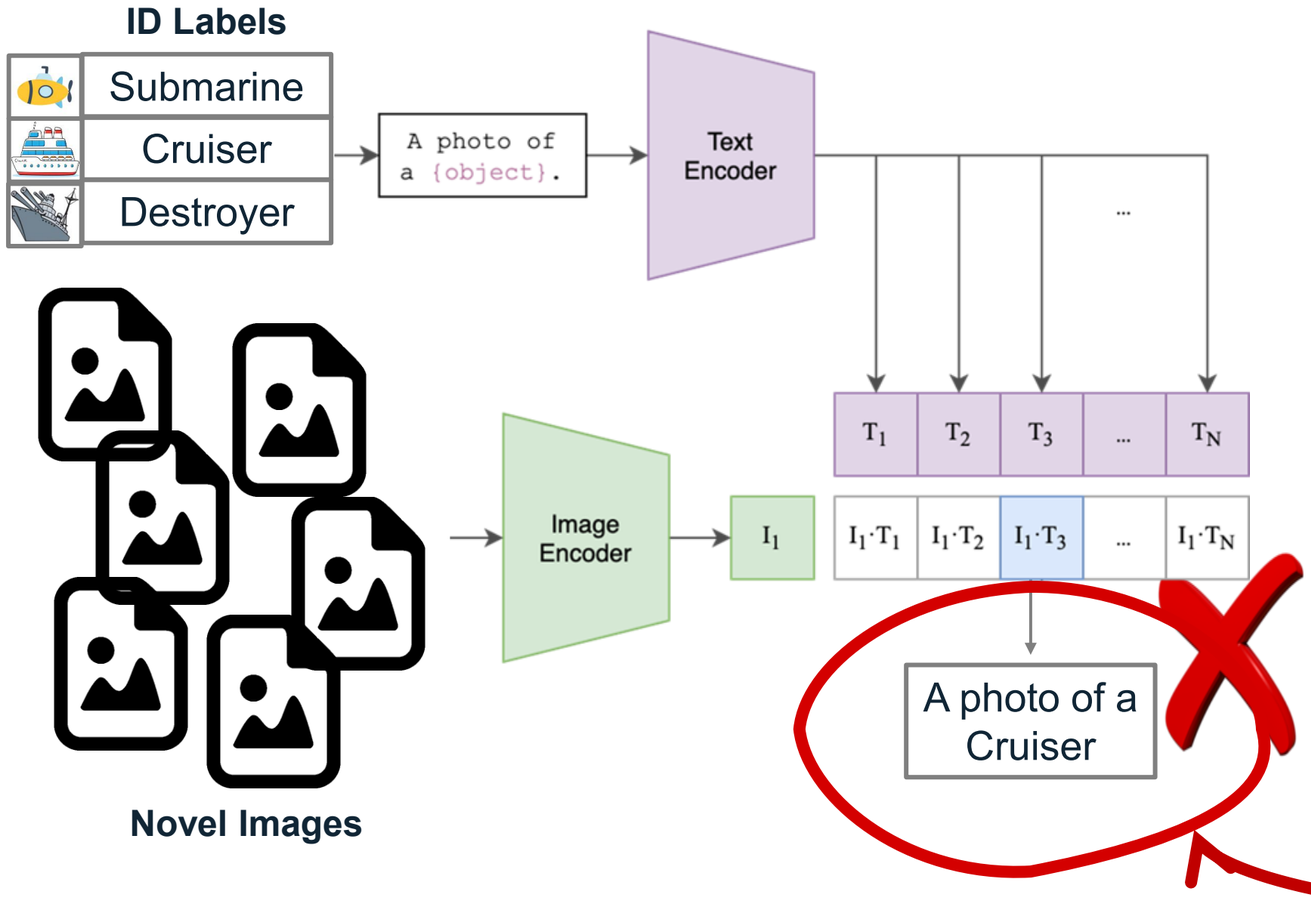


ID Classes

[0,0,1] [1,0,0] [0,1,0]

Out-of-Distribution (OOD)

Why Models Struggle with Novelty and Unexpected Data

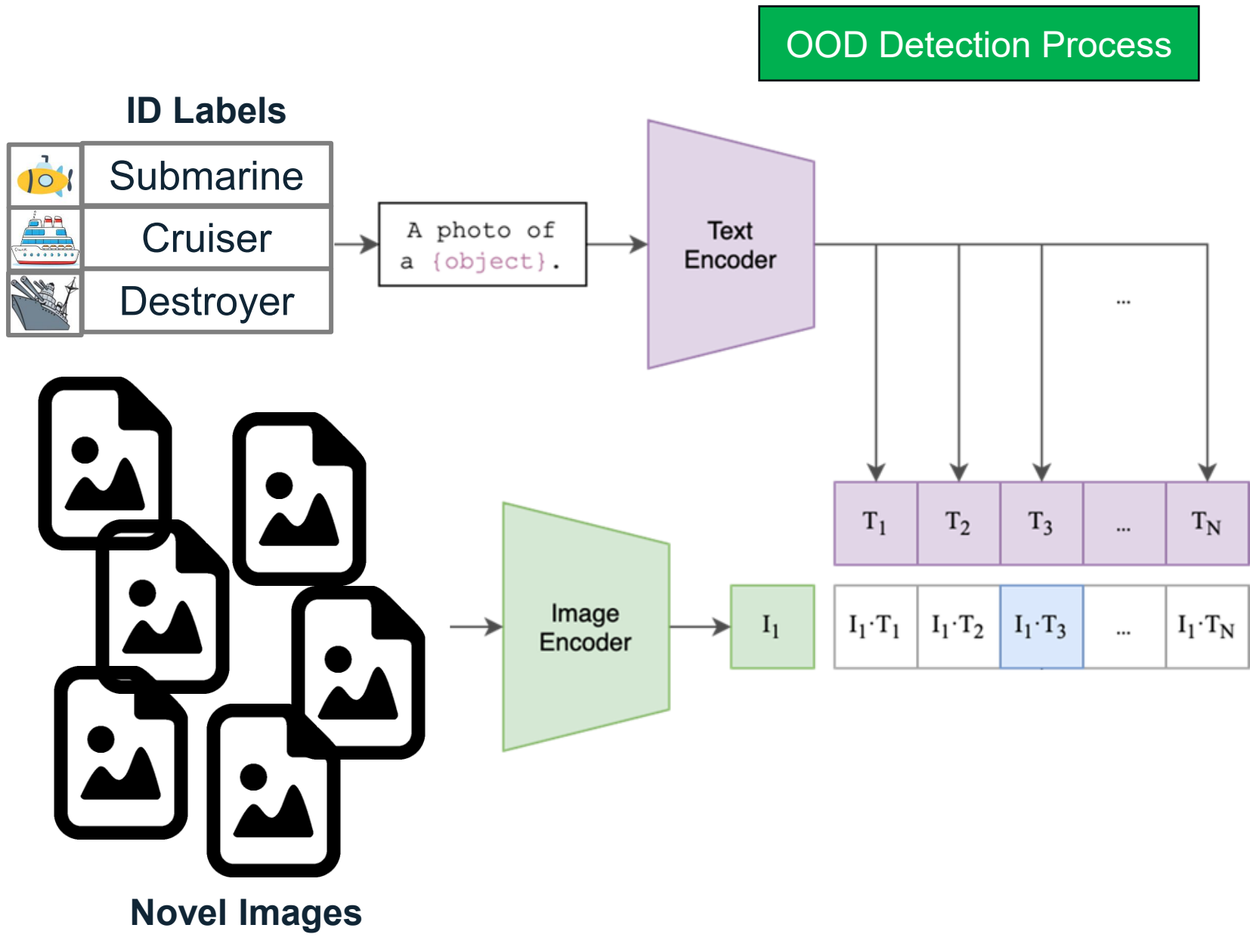


If no appropriate prompt is available, the model may incorrectly assign an OOD (Out-of-Distribution) **image to one of the ID** (In-Distribution) labels.

THE UNIVERSITY of ADELAIDE

Out-of-Distribution (OOD)

Why Models Struggle with Novelty and Unexpected Data



If this score exceeds a threshold (λ), the sample is considered **in-distribution**; otherwise, it is classified as **out-of-distribution**.

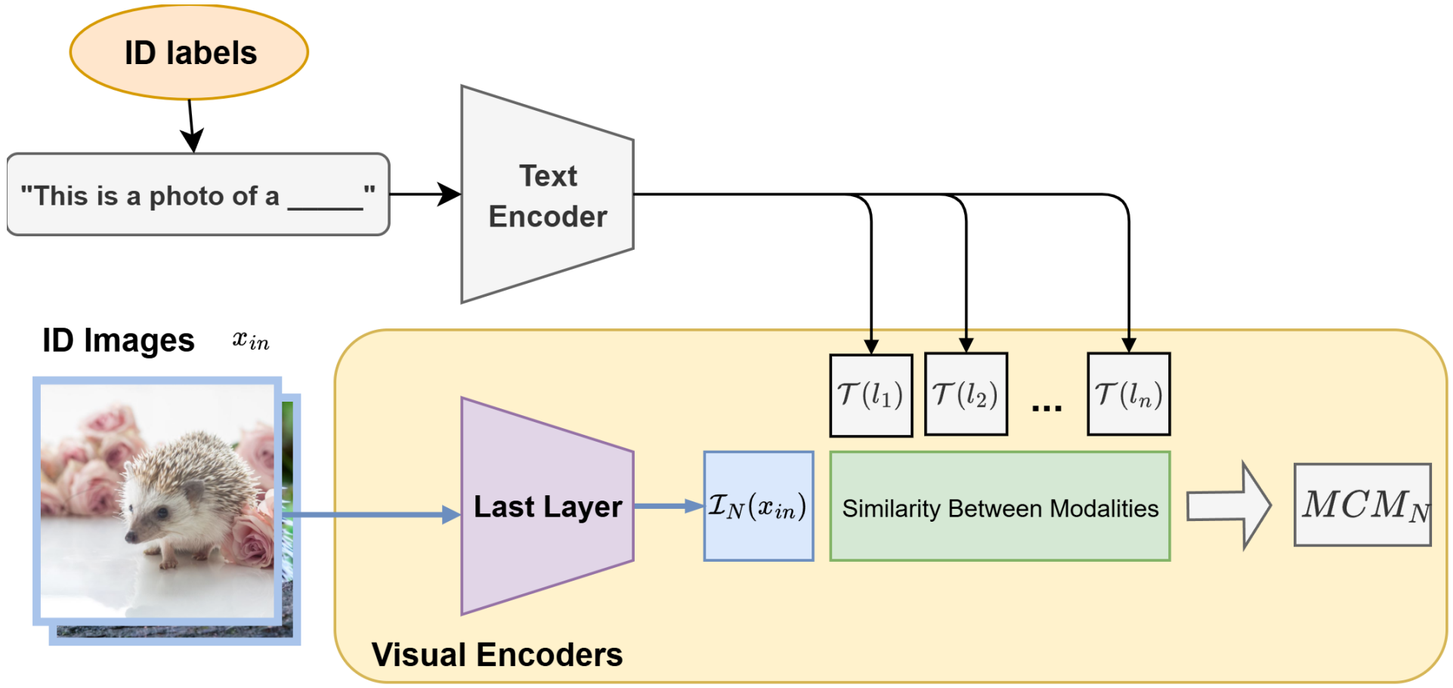
$$G_\lambda(\mathbf{x}; f) = \begin{cases} \text{in} & S(\mathbf{x}; f) \geq \lambda \\ \text{out} & S(\mathbf{x}; f) < \lambda \end{cases}$$

f: Depend on the model
S: Scoring Score

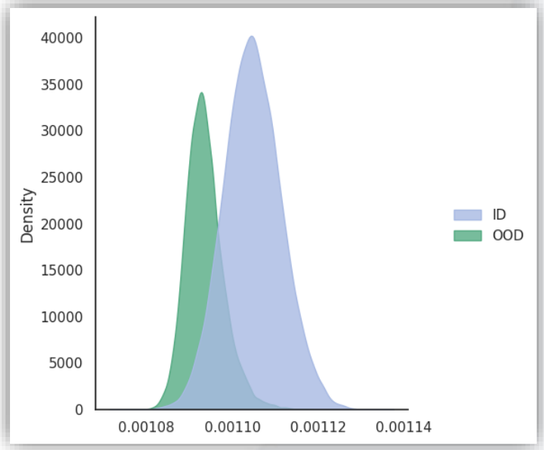
THE UNIVERSITY of ADELAIDE

Zero-shot OOD Detection

General way to attack the problem using vision-language representations



Maximum Concept Matching (MCM) is a **zero-shot classification method** that utilizes a model's existing structure for prediction without requiring dedicated fine-tuning.

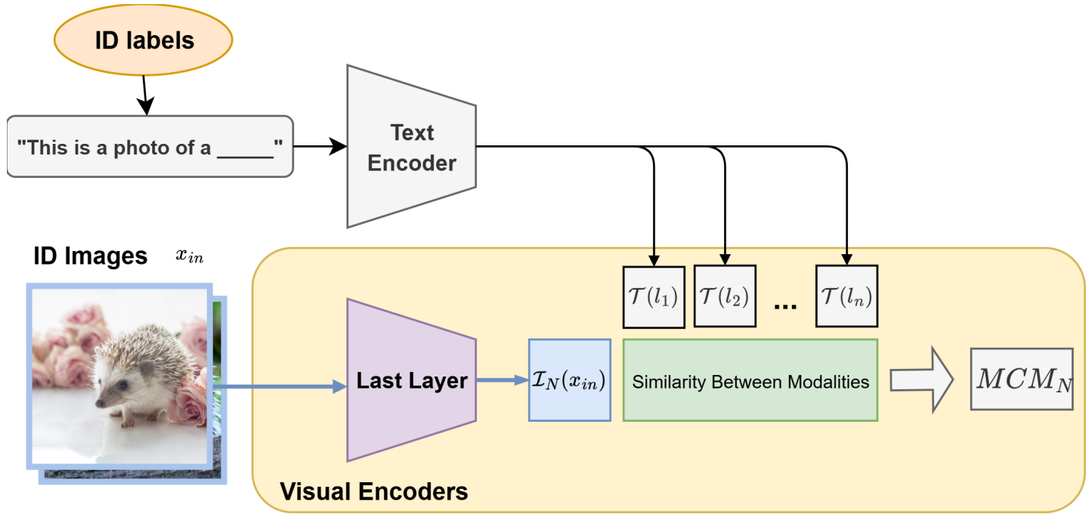


THE UNIVERSITY of ADELAIDE

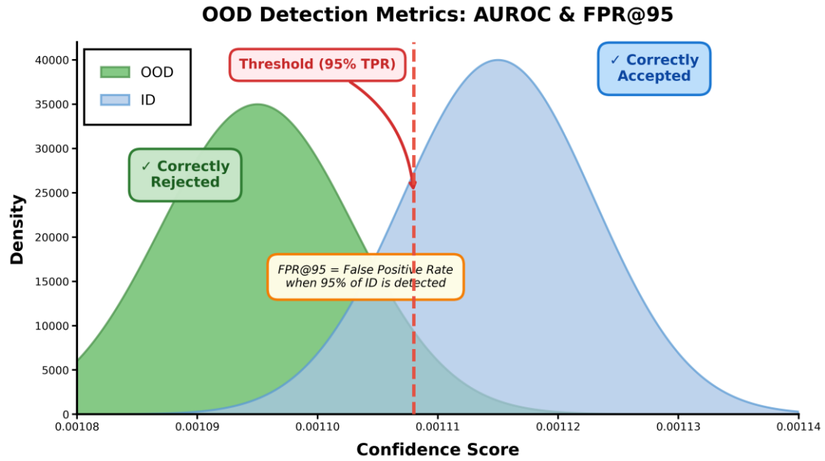
Zhou et al., *Delving into Out-of-Distribution Detection with Vision-Language Representations*, arXiv:2210.04250

Zero-shot OOD Detection

General way to attack the problem using vision-language representations



Maximum Concept Matching (MCM) is a **zero-shot classification method** that utilizes a model's existing structure for prediction without requiring dedicated fine-tuning.



AUROC: measures the overall separability between ID and OOD samples.

FPR@95: False positive rate when 95% of ID data is correctly detected.

THE UNIVERSITY of ADELAIDE

Zhou et al., *Delving into Out-of-Distribution Detection with Vision-Language Representations*, arXiv:2210.04250

THE UNIVERSITY of ADELAIDE

OOD Detection Using Intermediate Layers

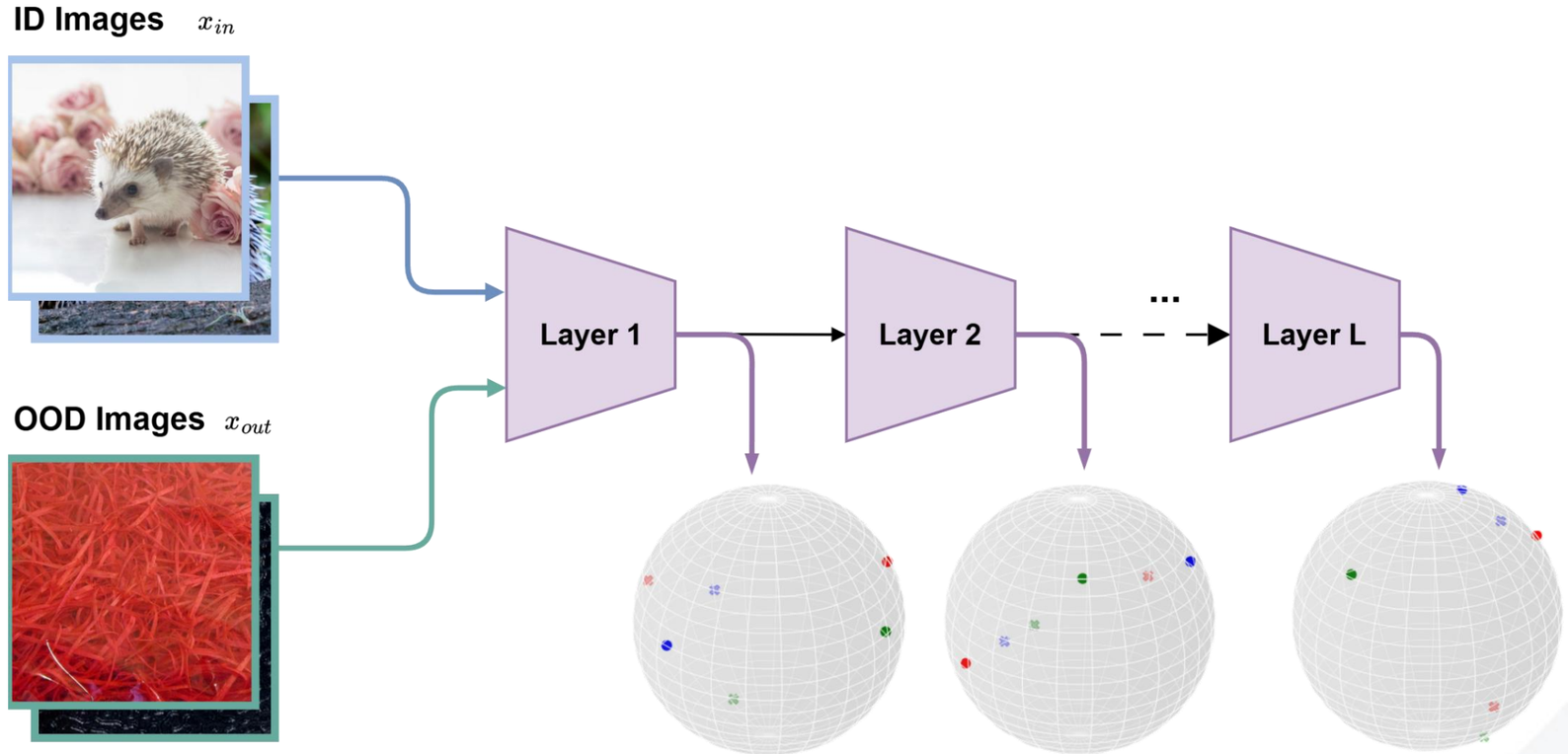
What role do intermediate layers play in out-of-distribution detection when using vision-language representations?

OOD Detection via Entropy-Guided Multi-Layer Fusion

Challenging a Core Assumption

Motivation: Exploiting Complementary Signals

- The final layer alone overlooks diverse and complementary signals encoded across intermediate representations.
- Fusing these signals exploits Representation Diversity and improves robustness by reducing prediction noise.

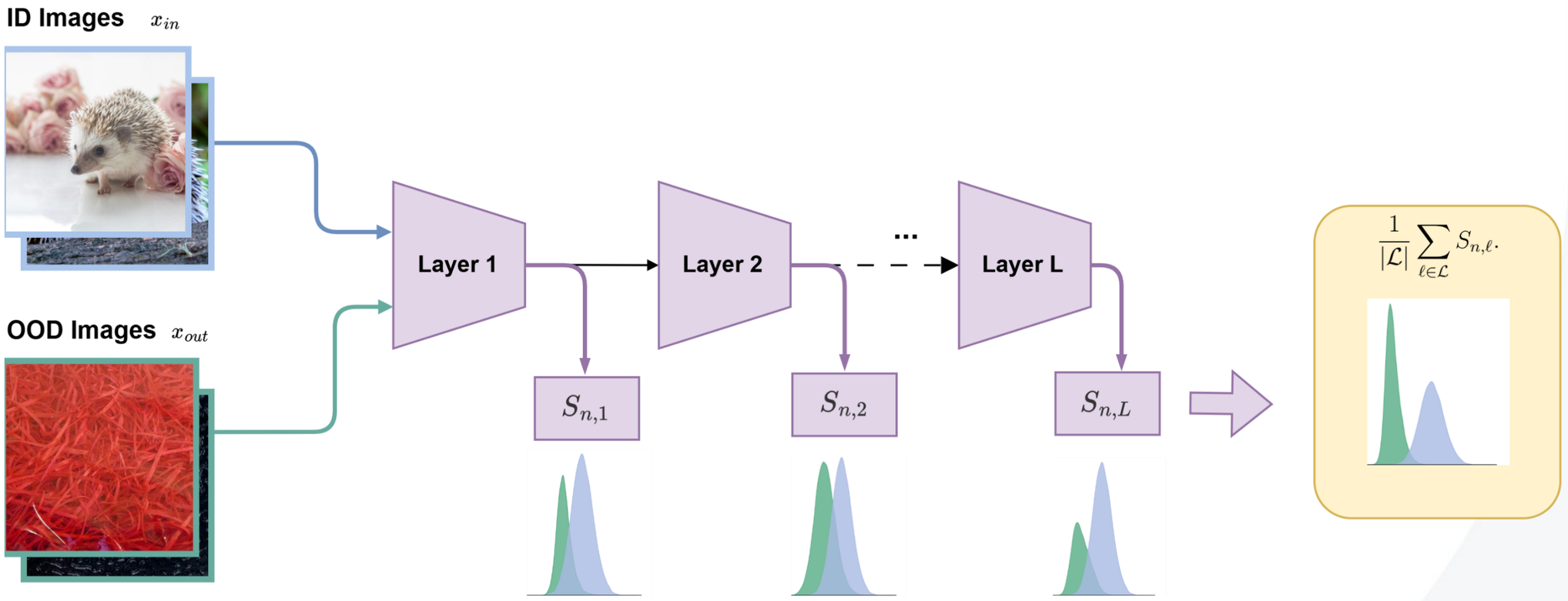


If predictions are obtained from each layer, the model produce distinct decisions at every layer.

THE UNIVERSITY of ADELAIDE

OOD Detection via Entropy-Guided Multi-Layer Fusion

Challenging a Core Assumption

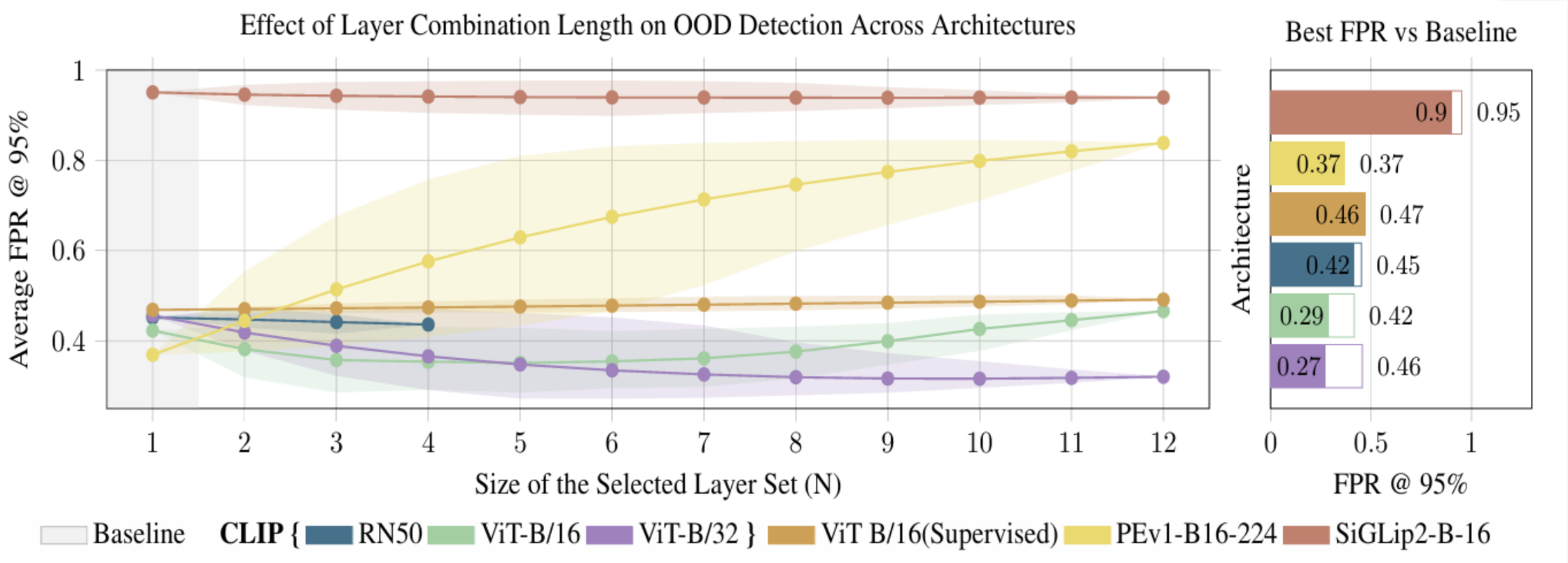


The objective is to select the layers that most effectively capture the ID, and by leveraging this complementary data, improve our capacity to differentiate between ID and OOD.

THE UNIVERSITY of ADELAIDE

The Power of Layer Fusion ⭐

Performance gains observed across multiple CLIP and vision backbones using just a single layer



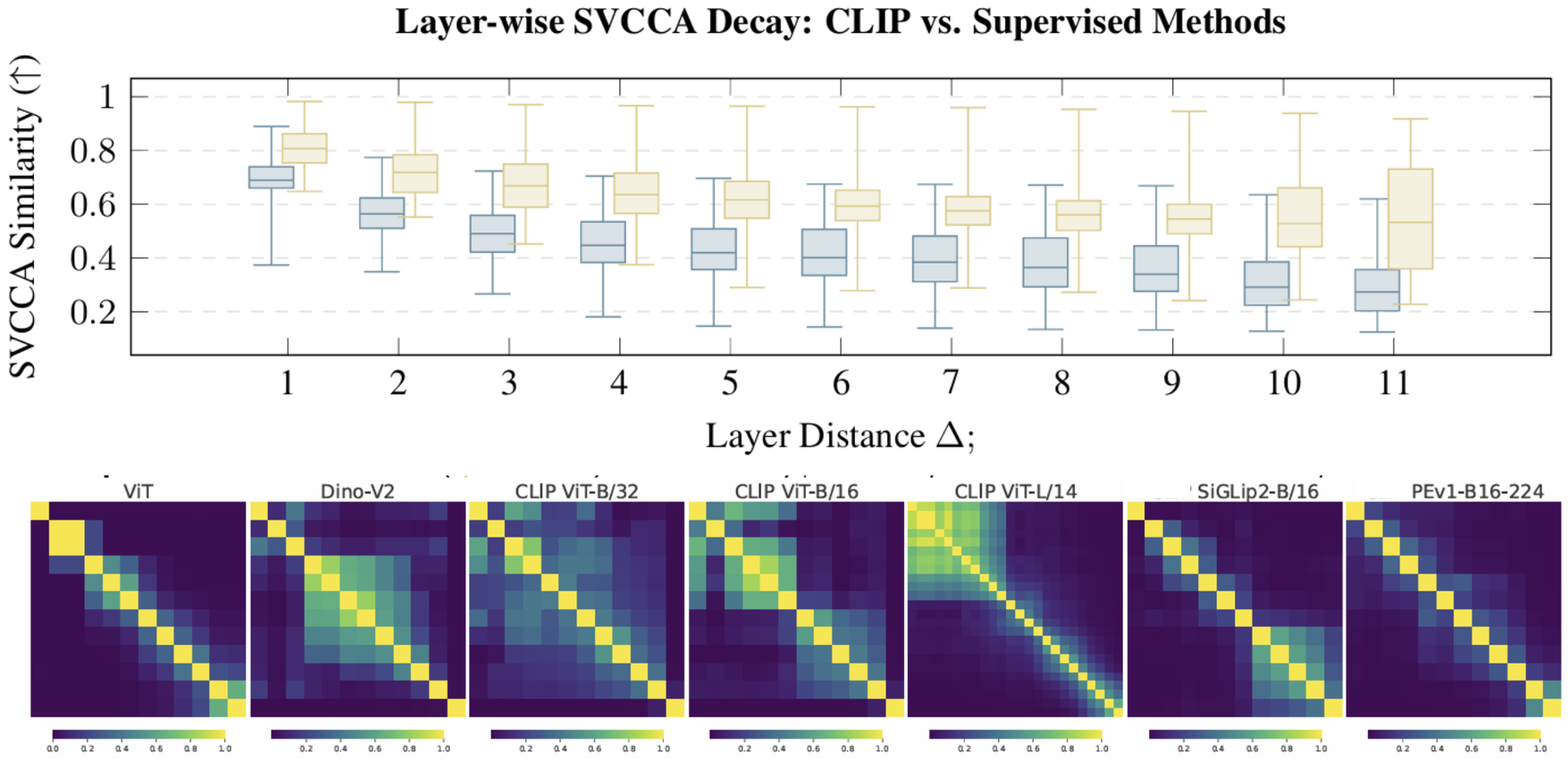
⭐ **Key Finding**

Our findings indicate that leveraging intermediate layers can enhance performance in out-of-distribution (OOD) detection, paving the way for a new direction in this area that typically relies solely on the final layer for decision-making.

THE UNIVERSITY of ADELAIDE

How Representations Evolve Across Models ⭐

Exploring How Early Layers Can Help Detect Out-of-Distribution Data



⭐ **Key Finding**

Architectures like CLIP, where layers are diverse yet consistent, benefit most from intermediate-layer fusion.
In contrast, models with unstable prediction profiles may suffer from interference without selective fusion.

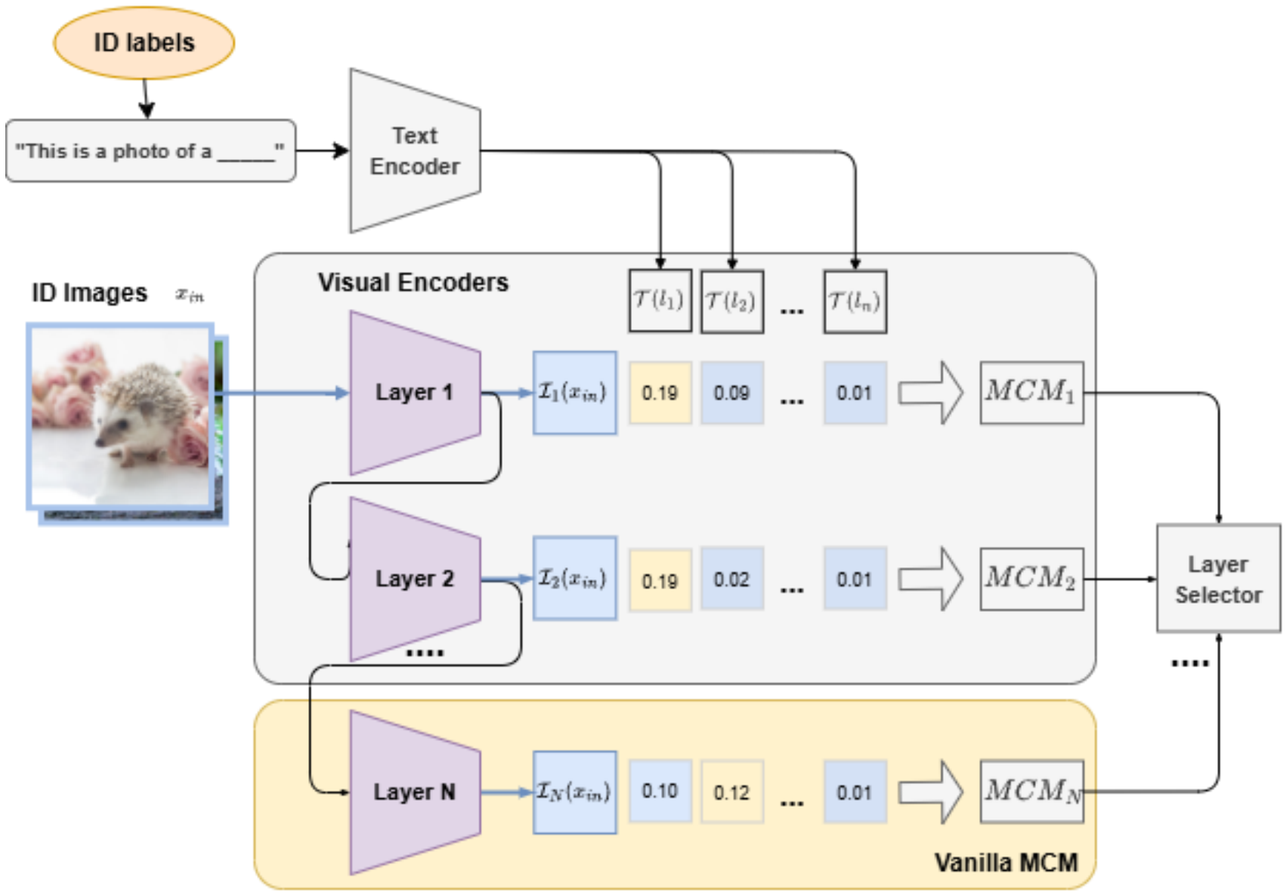
THE UNIVERSITY of ADELAIDE

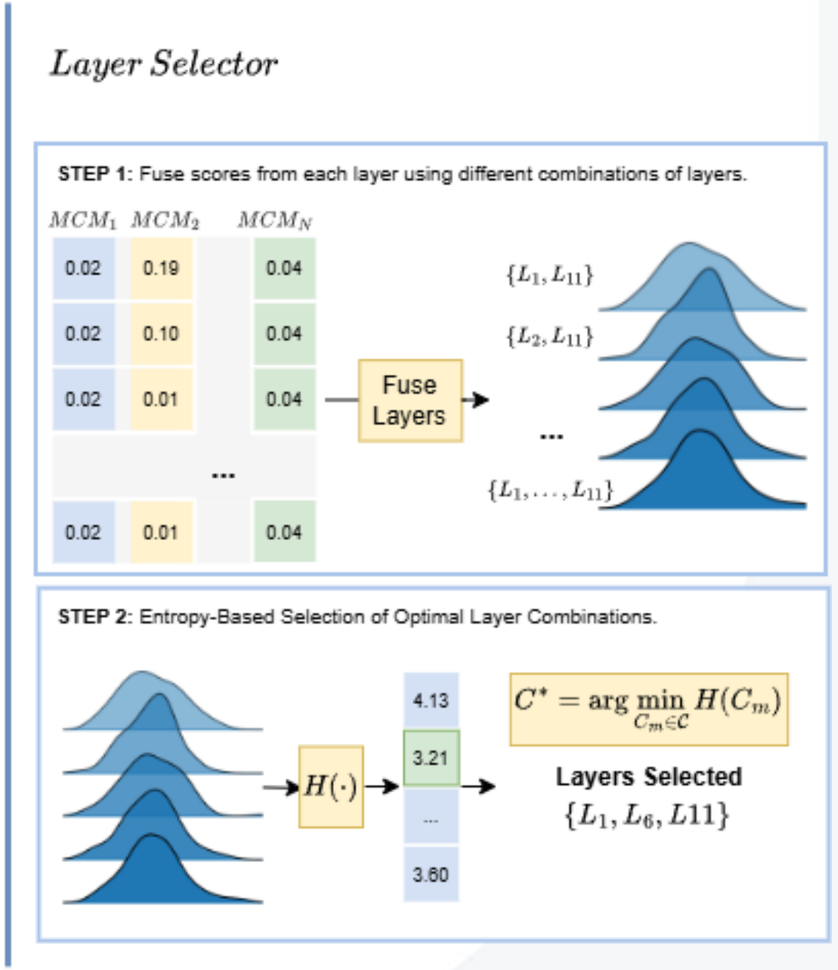
Leveraging these characteristics in CLIP

Utilizing intermediate layers within CLIP to enhance OOD performance

Our Approach ⚒️

Fusing Intermediate Representations for Robust Zero-Shot OOD Detection

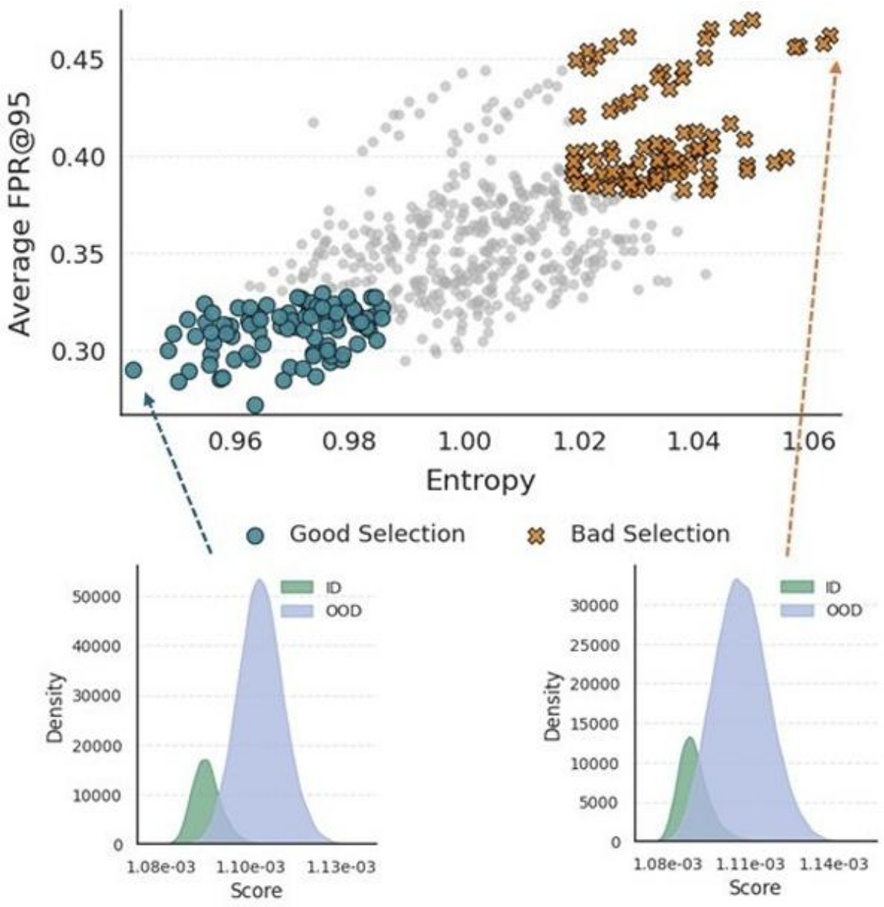




THE UNIVERSITY of ADELAIDE

Our Approach ⚒️

Fusing Intermediate Representations for Robust Zero-Shot OOD Detection



- **Layer permutations:** Each point in the plot corresponds to a different permutation of intermediate layers.
- **Entropy as selection criterion:** Using only ID data (without labels), we found a linear relation between the entropy of layer combinations and their OOD detection performance (FPR).

THE UNIVERSITY of ADELAIDE

Robust Gains Across Benchmarks ⭐

Consistent improvements in FPR and AUROC on diverse datasets and backbones

Method	Backbone	iNaturalist		SUN		Places		Texture		ImageNet22K		COCO		Average	
		FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑
ID - ImageNet1K															
MCM	ViT-B/32	33.85	93.62	40.99	91.56	46.71	89.25	60.90	85.03	-	-	-	-	45.61	89.87
MCM	ViT-B/16	30.67	94.63	37.41	92.57	43.67	89.96	57.34	86.18	-	-	-	-	42.27	90.83
SeTAR + MCM	ViT-B/16	26.92	94.67	35.57	92.79	42.64	90.16	55.83	86.58	-	-	-	-	40.24	91.05
GL-MCM	ViT-B/16	17.42	96.44	30.75	**93.44**	37.62	90.63	55.20	85.54	-	-	-	-	35.25	91.51
SeTAR + GL-MCM	ViT-B/16	13.36	96.92	**28.17**	93.36	36.80	90.40	54.17	84.59	-	-	-	-	33.12	91.32
Ours	ViT-B/16	15.98	96.90	45.58	89.69	35.71	**92.72**	**25.51**	**94.84**	-	-	-	-	30.70	**93.54**
Ours	ViT-B/32	**12.02**	**97.64**	28.98	**93.37**	**35.69**	91.61	39.36	91.07	-	-	-	-	**29.01**	93.42
ID - Pascal-VOC															
MCM	ViT-B/32	34.80	95.35	30.60	93.74	37.70	91.99	51.60	91.68	55.00	91.16	59.10	89.23	44.80	92.19
MCM	ViT-B/16	10.51	97.93	30.45	94.25	36.11	91.86	53.21	91.77	53.82	91.12	57.10	89.02	40.20	92.66
SeTAR + MCM	ViT-B/16	4.38	98.70	26.24	94.95	28.67	93.28	50.32	92.32	44.61	92.63	49.80	89.68	34.00	93.59
GL-MCM	ViT-B/16	4.33	98.81	22.94	94.63	26.20	93.11	41.61	92.88	37.88	93.17	43.70	90.71	29.44	93.88
SeTAR + GL-MCM	ViT-B/16	3.01	**99.04**	21.76	94.98	24.00	93.73	37.61	93.87	**33.46**	**94.24**	**40.60**	**91.48**	**26.74**	**94.56**
Ours	ViT-B/32	32.18	94.40	35.60	92.61	49.17	92.34	**33.72**	**95.78**	54.36	91.34	57.60	89.97	43.77	92.74
Ours	ViT-B/16	**2.19**	98.92	**19.70**	**95.77**	**19.53**	**95.11**	44.08	92.06	34.70	91.15	41.60	88.09	26.97	93.52

THE UNIVERSITY of ADELAIDE

Robust Gains Across Benchmarks ⭐

Consistent improvements in FPR and AUROC on diverse datasets and backbones

Method	Backbone	iNaturalist		SUN		Places		Texture		ImageNet22K		COCO		Average	
		FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑
ID - ImageNet1K															
MCM	ViT-B/32	33.85	93.62	40.99	91.56	46.71	89.25	60.90	85.03	-	-	-	-	45.61	89.87
MCM	ViT-B/16	30.67	94.63	37.41	92.57	43.67	89.96	57.34	86.18	-	-	-	-	42.27	90.83
SeTAR + MCM	ViT-B/16	26.92	94.67	35.57	92.79	42.64	90.16	55.83	86.58	-	-	-	-	40.24	91.05
GL-MCM	ViT-B/16	17.42	96.44	30.75	**93.44**	37.62	90.63	55.20	85.54	-	-	-	-	35.25	91.51
SeTAR + GL-MCM	ViT-B/16	13.36	96.92	**28.17**	93.36	36.80	90.40	54.17	84.59	-	-	-	-	33.12	91.32
Ours	ViT-B/16	15.98	96.90	45.58	89.69	35.71	**92.72**	**25.51**	**94.84**	-	-	-	-	30.70	**93.54**
Ours	ViT-B/32	**12.02**	**97.64**	28.98	**93.37**	**35.69**	91.61	39.36	91.07	-	-	-	-	**29.01**	93.42
ID - Pascal-VOC															
MCM	ViT-B/32	34.80	95.35	30.60	93.74	37.70	91.99	51.60	91.68	55.00	91.16	59.10	89.23	44.80	92.19
MCM	ViT-B/16	10.51	97.93	30.45	94.25	36.11	91.86	53.21	91.77	53.82	91.12	57.10	89.02	40.20	92.66
SeTAR + MCM	ViT-B/16	4.38	98.70	26.24	94.95	28.67	93.28	50.32	92.32	44.61	92.63	49.80	89.68	34.00	93.59
GL-MCM	ViT-B/16	4.33	98.81	22.94	94.63	26.20	93.11	41.61	92.88	37.88	93.17	43.70	90.71	29.44	93.88
SeTAR + GL-MCM	ViT-B/16	3.01	**99.04**	21.76	94.98	24.00	93.73	37.61	93.87	**33.46**	**94.24**	**40.60**	**91.48**	**26.74**	**94.56**
Ours	ViT-B/32	32.18	94.40	35.60	92.61	49.17	92.34	**33.72**	**95.78**	54.36	91.34	57.60	89.97	43.77	92.74
Ours	ViT-B/16	**2.19**	98.92	**19.70**	**95.77**	**19.53**	**95.11**	44.08	92.06	34.70	91.15	41.60	88.09	26.97	93.52

Our method consistently outperforms the baseline MCM across datasets, reducing false positive rates

THE UNIVERSITY of ADELAIDE

Robust Gains Across Benchmarks ⭐

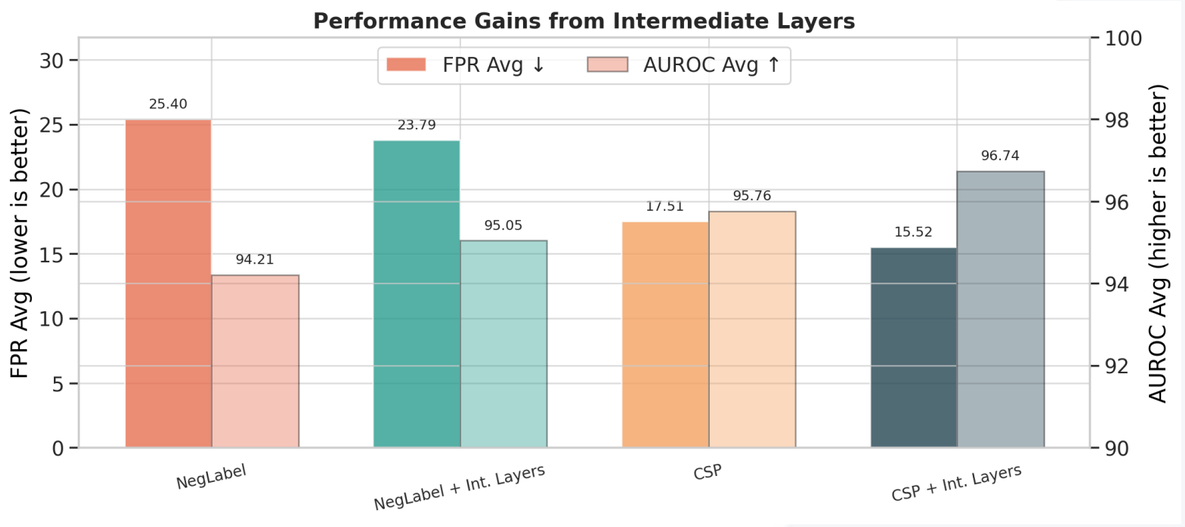
Consistent improvements in FPR and AUROC on diverse datasets and backbones

Table 4: **Intermediate vs last layer performance improvement across architectures.** We report AUC improvement (%, ↑) and FPR improvement (%, ↑ indicates lower FPR) for uncertainty quantification methods. Results averaged across SUN, Places365, DTD, and iNaturalist datasets.

Method	Type	ResNet		ViT-B/32		ViT-B/16		Average	
		AUC↑	FPR↑	AUC↑	FPR↑	AUC↑	FPR↑	AUC↑	FPR↑
Distributional Methods									
Entropy	Int vs Last	+0.1	+0.0	+1.1	+2.5	+9.0	+16.0	+3.4	+6.2
Energy	Int vs Last	+0.5	+0.2	+4.2	+0.4	+2.4	+0.6	+2.4	+0.4
Variance	Int vs Last	+0.1	+0.1	+1.0	+2.3	+8.9	+16.0	+3.3	+6.1
Max-based Methods									
MaxLogit	Int vs Last	+1.2	+4.0	-0.9	+0.3	+3.2	+8.6	+1.2	+4.3
MCM	Int vs Last	+0.8	+3.5	+3.6	+16.6	+2.7	+11.6	+2.4	+10.2

Our method enhances both **distributional** and **max-based** scoring strategies, showing consistent gains in AUROC and reductions in FPR.

When applied to strong baselines like **NegLabel*** and **CSP****, our approach further lowers false positives and improves robustness, confirming its complementarity with existing OOD detection methods.



*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."
**Mengyuan Chen, Junyu Gao, and Changsheng Xu. "Conjugated Semantic Pool Improves OOD Detection with Pre-trained Vision-Language Models."

THE UNIVERSITY of ADELAIDE

Summary ⭐

- This study explores the impact of intermediate layers in different models on the task of out-of-distribution (OOD) detection.
- Leveraging these intermediate layers can significantly enhance the models' ability to identify OOD samples.
- We propose a baseline criterion for selecting layers without requiring training, using a zero-shot approach.
- This technique can be applied to various methods that utilize CLIP as their backbone.

THE UNIVERSITY of ADELAIDE

make history.

THE UNIVERSITY of ADELAIDE