SeTAR: Out-of-Distribution Detection with Selective Low-Rank Approximation

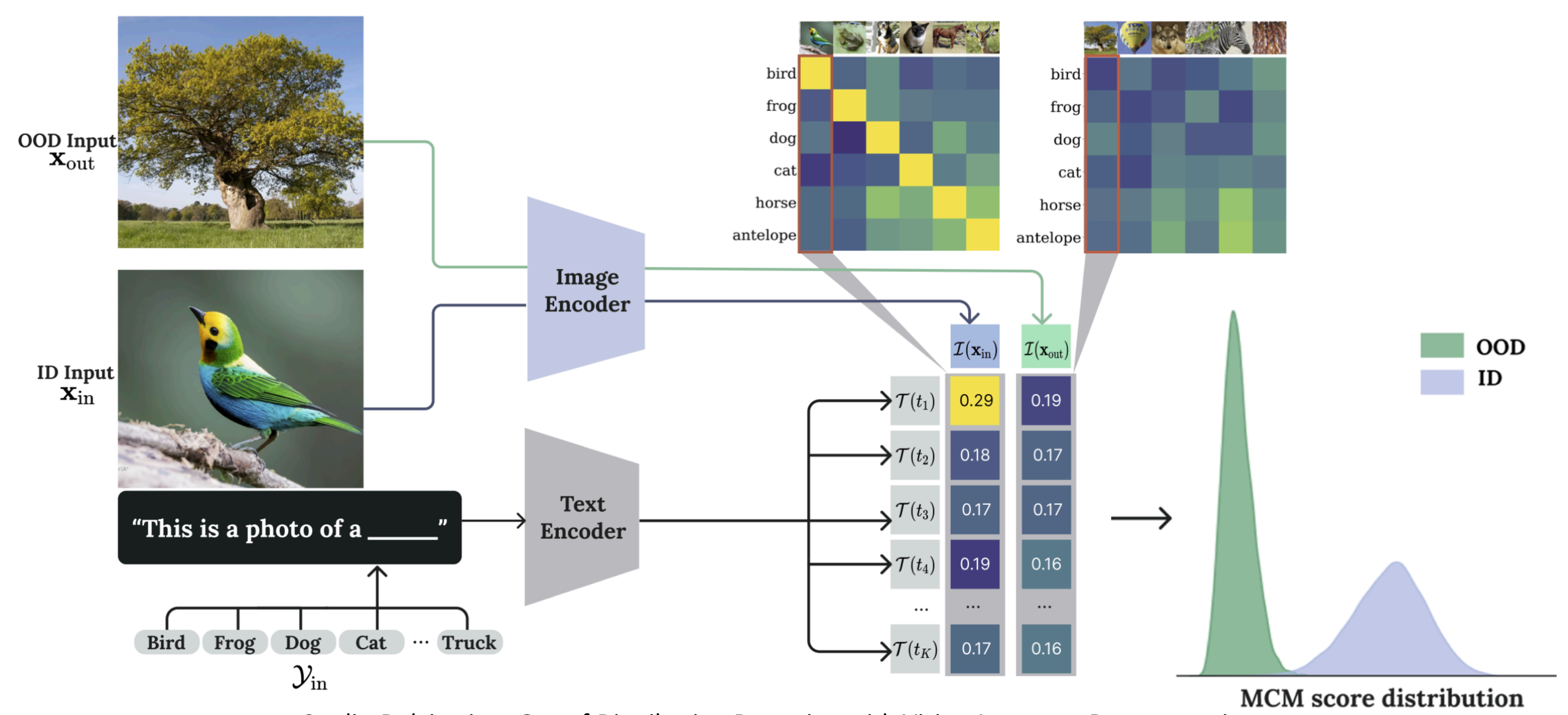
Yixia Li[1*] Boya Xiong[2*] Guanhua Chen[1†] Yun Chen[2†]

[1]Southern University of Science and Technology [2]Shanghai University of Finance Economics

[*]Equal Contribution [†]Corresponding Authors

Paper Code Project Homepage

CLIP-based Zero-shot OOD Detection



Credit: Delving into Out-of-Distribution Detection with Vision-Language Representations

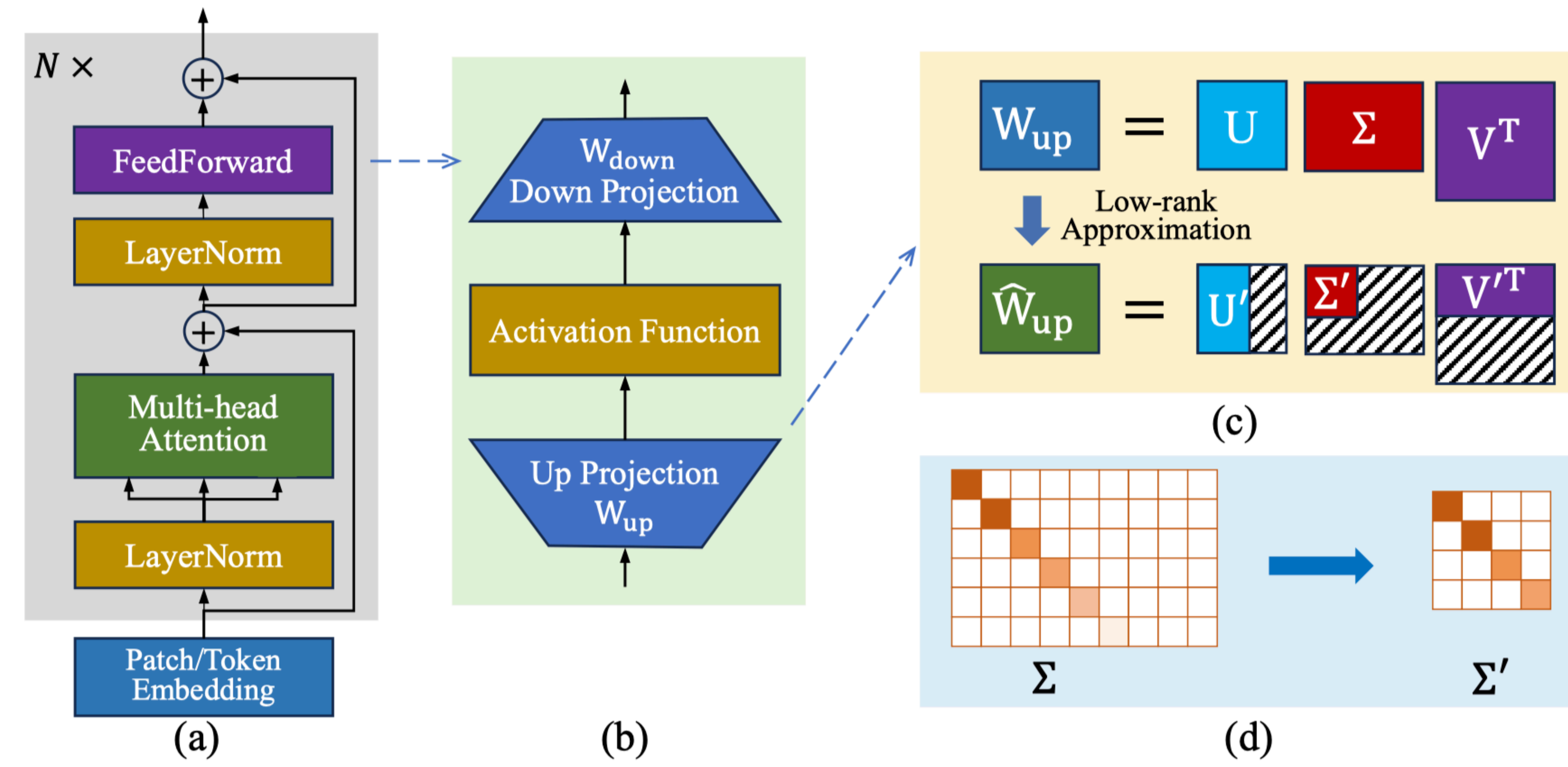
Background & Motivation

- OOD detection is crucial for real-world applications like healthcare and autonomous vehicles. Traditional methods lack generalizability, especially in large, pre-trained models like CLIP.
- Existing CLIP-based OOD methods require complex fine-tuning or underperform in zero-shot settings due to domain mismatch, highlighting the need for a training-free approach that preserves model integrity.

Research Gap

- Zero-shot methods: Limited by domain gaps, reducing performance.
- Fine-tuning methods: Risk disrupting pretrained representations.
- CNN post-hoc methods: Rely on distinct activation patterns for ID/OOD data, which doesn't applied to large-scale pre-trained models such as CLIP.

SeTAR: Selective Low-Rank Approximation



Key Contribution

- A training-free OOD detection approach using SVD-based Selective Low-Rank Approximation to enhance performance without additional training.
- Achieves SOTA results on ImageNet1K and Pascal-VOC benchmarks, outperforming existing zero-shot and fine-tuning methods.

SeTAR Overview

- Low-Rank Approximation: Retains critical model weights by discarding minor singular components.
- Greedy Search: Optimizes rank reduction layer-by-layer, targeting linear matrix Wup for maximum impact.
- SeTAR+FT: A fine-tuning extension, freezes major components and tuning minor ones, enabling parameters-efficient fine-tuning.

Key Findings

1. Training-Free (SeTAR): Achieves state-of-the-art zero-shot OOD detection across different backbones (CLIP-base, CLIP-large, Swin-base) and various score functions (MCM, GL-MCM, MSP, Energy) on ImageNet1K and VOC.
2. Fine-Tuning (SeTAR+FT): Outperforms existing fine-tuning methods (LoCoOp, LoRA) with faster convergence, superior OOD detection, and higher classification accuracy, especially on challenging datasets.
3. Image Classification Accuracy: Maintains or further enhances image classification performance while strengthening OOD detection capabilities.
4. Near-OOD Effectiveness: SeTAR and SeTAR+FT perform well even on challenging near-OOD tasks, confirming robustness.
5. Pruning Strategies: Discarding minor singular values optimally enhances OOD detection by preserving critical components.
6. Modality Impact: Vision modality yields the best gains; combining vision and text maximizes results.

1. Training-Free Results (SeTAR)

Backbone	Score	Vanilla Method		SeTAR	
		FPR↓	AUC↑	FPR↓	AUC↑
ImageNet1K					
CLIP-base	MCM	43.09	90.74	**40.24**	**91.05**
CLIP-base	GL-MCM	35.29	90.86	**33.12**	**91.32**
CLIP-large	MCM	37.19	91.73	**36.26**	**91.92**
CLIP-large	GL-MCM	40.65	89.98	**39.54**	**90.22**
Swin-base	MSP	59.25	84.12	**56.05**	**85.77**
Swin-base	Energy	65.01	76.10	**51.61**	**84.42**
Pascal-VOC					
CLIP-base	MCM	37.24	92.98	**32.46**	**93.74**
CLIP-base	GL-MCM	29.44	93.88	**23.86**	**94.87**
CLIP-large	MCM	52.21	91.68	**42.57**	**92.91**
CLIP-large	GL-MCM	43.96	92.45	**31.12**	**94.00**

2. Fine-tuning Results (SeTAR+FT)

CLIP-base	MCM Score		GL-MCM Score	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑
NPOS†	42.20	90.43	36.86	90.37
CoOp†	44.81	90.03	36.58	90.25
LoCoOp†	40.17	91.53	33.52	92.14
LoCoOp*	$39.76_{\pm 4.06}$	$91.22_{\pm 0.52}$	$34.14_{\pm 1.64}$	$91.73_{\pm 0.17}$
LoRA*	$41.67_{\pm 0.14}$	$90.85_{\pm 0.01}$	$34.36_{\pm 0.11}$	$90.88_{\pm 0.01}$
SeTAR+FT	$\mathbf{38.77}_{\pm 0.22}$	$\mathbf{91.55}_{\pm 0.01}$	$\mathbf{32.19}_{\pm 0.20}$	$\mathbf{92.31}_{\pm 0.05}$
CLIP-large	MCM Score		GL-MCM Score	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑
LoCoOp*	$40.74_{\pm 3.80}$	$91.13_{\pm 0.79}$	$46.74_{\pm 4.19}$	$89.32_{\pm 0.80}$
LoRA*	$38.62_{\pm 0.07}$	$91.66_{\pm 0.02}$	$43.39_{\pm 0.01}$	$89.76_{\pm 0.03}$
SeTAR+FT	$\mathbf{34.75}_{\pm 0.55}$	$\mathbf{92.86}_{\pm 0.15}$	$\mathbf{37.05}_{\pm 0.59}$	$\mathbf{91.83}_{\pm 0.12}$
Swin-base	MSP Score		Energy Score	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑
LoRA*	$57.02_{\pm 0.03}$	$80.49_{\pm 0.01}$	$62.17_{\pm 0.02}$	$72.80_{\pm 0.00}$
SeTAR+FT	$\mathbf{47.12}_{\pm 0.42}$	$\mathbf{87.80}_{\pm 0.44}$	$\mathbf{39.29}_{\pm 0.57}$	$\mathbf{88.01}_{\pm 0.51}$

3. Image Classification Accuracy

Method	IN1K	SUN	Places	Texture	**Average**
Vanilla CLIP*	64.07	75.77	45.65	43.60	57.27
LoCoOp*	64.93	75.89	46.47	37.79	56.27
LoRA*	65.43	76.86	46.58	**43.98**	58.21
SeTAR	63.97	75.50	45.81	43.76	57.26
SeTAR+FT	**67.02**	**77.94**	**46.64**	43.28	**58.72**

4. Near-OOD Results

Method	Category	MCM Score		GL-MCM Score	
		FPR↓	AUC↑	FPR↓	AUC↑
Vanilla	Training-Free	89.28	63.88	85.62	67.63
SeTAR	Training-Free	88.29	64.20	**84.03**	68.29
LoCoOp	Training-Free	89.72	63.45	86.79	65.93
LoRA	Finetuning	88.52	65.38	84.39	68.85
SeTAR+FT	Finetuning	**87.16**	**68.13**	84.72	**70.42**

5. Pruning Strategies

Score	Principle		Random		Minor	
	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑
ImageNet1K						
MCM	43.09	90.74	43.09	90.74	**40.24**	**91.05**
GL-MCM	35.29	90.86	35.29	90.86	**33.12**	**91.32**
Pascal-VOC						
MCM	38.20	92.44	33.57	93.09	**32.46**	**93.74**
GL-MCM	25.36	93.67	26.20	94.66	**23.86**	**94.87**

6. Modality Impact

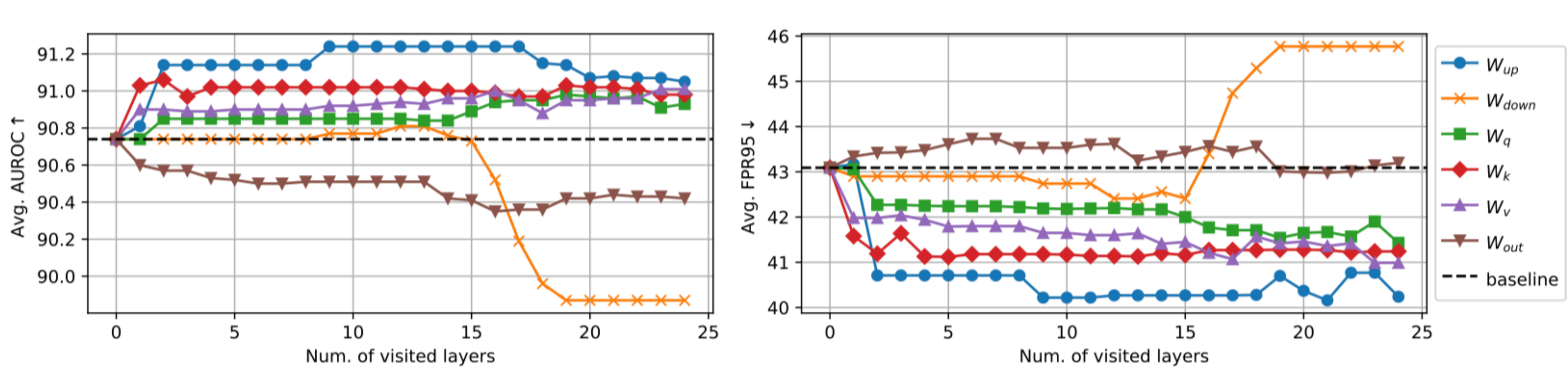
Score	Vision		Text		Vision+Text	
	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑
ImageNet1K						
MCM	40.27	**91.24**	42.78	90.50	**40.24**	91.05
GL-MCM	**32.97**	**91.60**	35.82	90.55	33.12	91.32
Pascal-VOC						
MCM	33.19	93.45	33.47	93.42	**32.46**	**93.74**
GL-MCM	24.88	94.51	24.59	94.52	**23.86**	**94.87**

Greedy Search Algorithm

Develop a rank reduction ratio for each layer to optimize detection performance.

1. Enumerate all vision encoder layers and then text encoder layers.
2. Apply different rank reduction ratio to the linear matrix.
3. Select the rank reduction ratio candidate with minimum loss.

- Wup is the most effective matrix across all linear weight matrices.



Loss Function

Pushes OOD samples far from ID samples while keeping ID samples close.

- Key Challenges: OOD images are unavailable during searching.
- Main Idea: Create pseudo OOD features with ID-irrelevant nuisances (e.g., backgrounds) in CLIP's local features.