Multi-Method Ensemble for Out-of-Distribution Detection

Lucas Rakotoarivony

Thales Research and Technology, cortAIx Labs,
1 Av. Augustin Fresnel, 91120, Palaiseau, France

BMVC 2025

1. Out-of-Distribution (OOD) detection

Neural Networks (NN) and OOD

- ▷ Can only recognize objects seen during training
- ▷ May produce overconfident predictions on unknown objects
- ▷ Need OOD detectors to distinguish known from unknown

Two families of OOD detectors

- ▷ **Scoring function** : produce high scores for ID samples from NN outputs
- ▷ **Feature truncation** : enhance separation between ID and OOD samples by modifying propagated signals

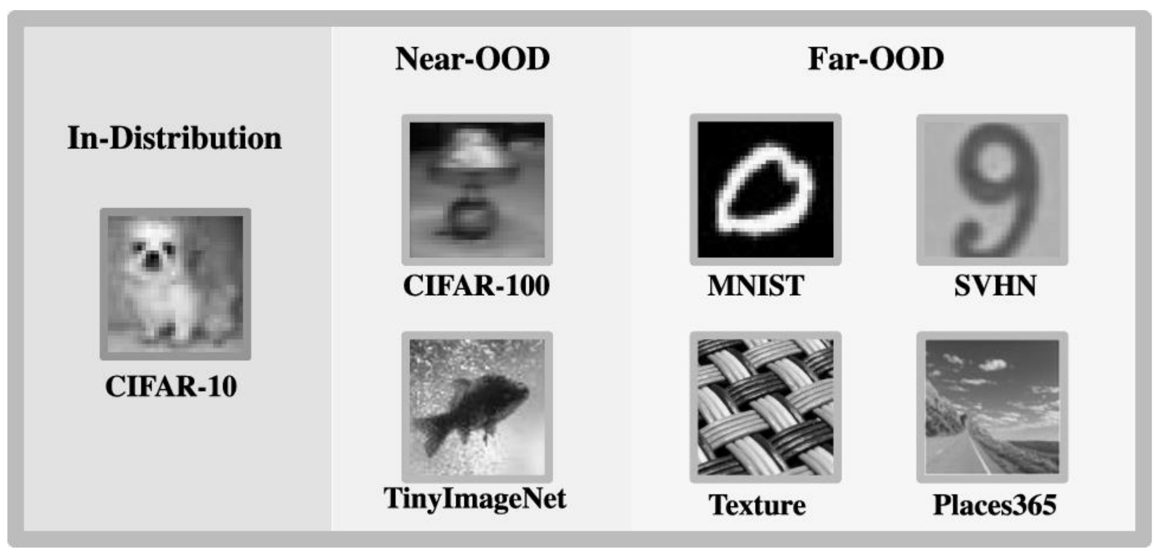
OOD main challenge: recognizing OOD samples that vary from slightly different to entirely unrelated

- ▷ Near-OOD
  - Similar images, different class
- ▷ Far-OOD
  - Visually and semantically distant

Scoring function equation for OOD detection

$$\mathbf{x} \in \begin{cases} \text{ID} & \text{if} \quad S(\mathbf{x}) \geq \tau \\ \text{OOD} & \text{if} \quad S(\mathbf{x}) < \tau \end{cases}$$

Illustration of OOD sample variability



[1]

Problematic : Does combining multiple scoring functions and feature truncations improve robustness across OOD types?

Our contributions :

- ▷ **Theory** : Proof that merging multiple OOD detectors enhances OOD detection
- ▷ **Methodology** : Proposal of a post-hoc strategy named Multi-Method Ensemble (MME)
- ▷ **Performance** : Benchmarking demonstrating MME's superiority over existing methods

2. Ensembling strategy for OOD detection

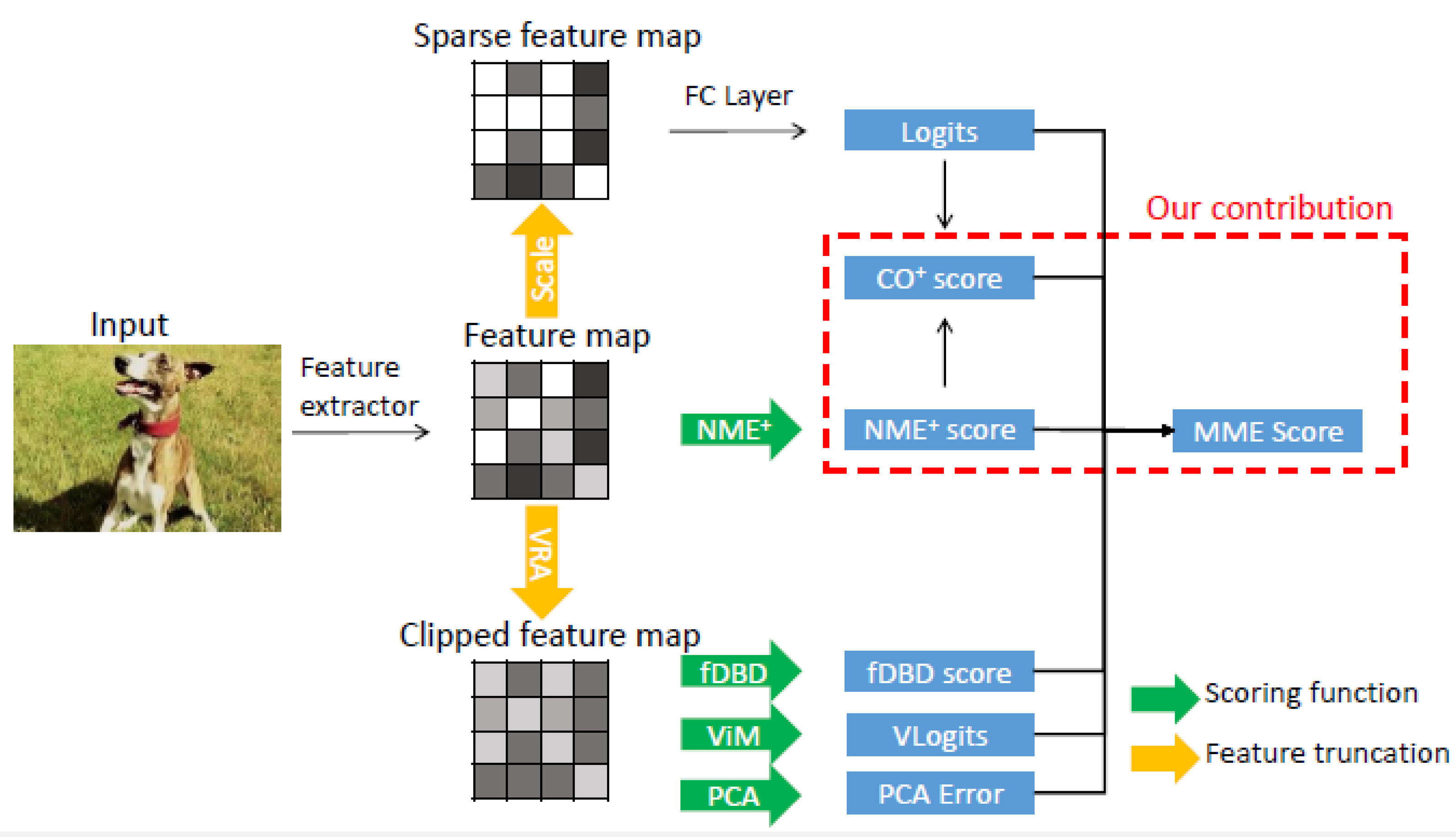
Properties that improve OOD detection

- ▷ Combining multiple scoring functions
  - Theoretical proof
  - Insight of literature
- ▷ Pairing feature truncation with scoring function
  - Empirical analysis
  - Insight of literature

MME framework overview

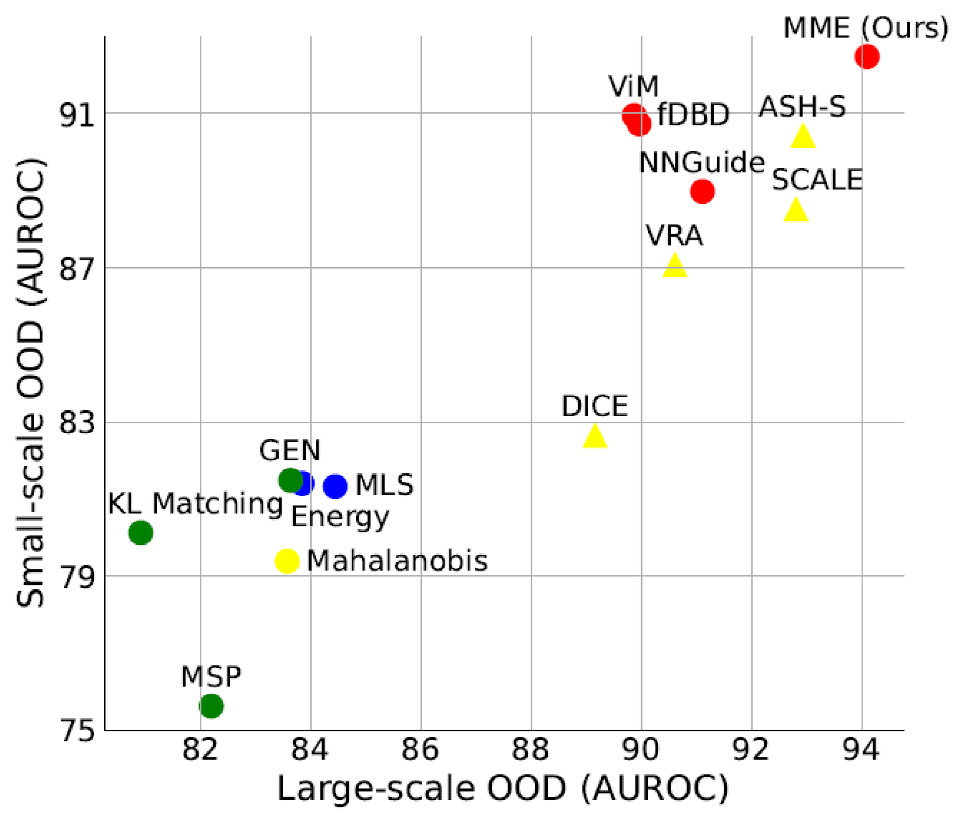
- ▷ MME implementation features:
  - 6 scoring functions
  - 2 feature truncations
- ▷ Benchmarked across diverse OOD scenarios:
  - small- vs large-scale OOD
  - near- vs far-OOD
- ▷ Evaluated against 16 state-of-the-art OOD detectors

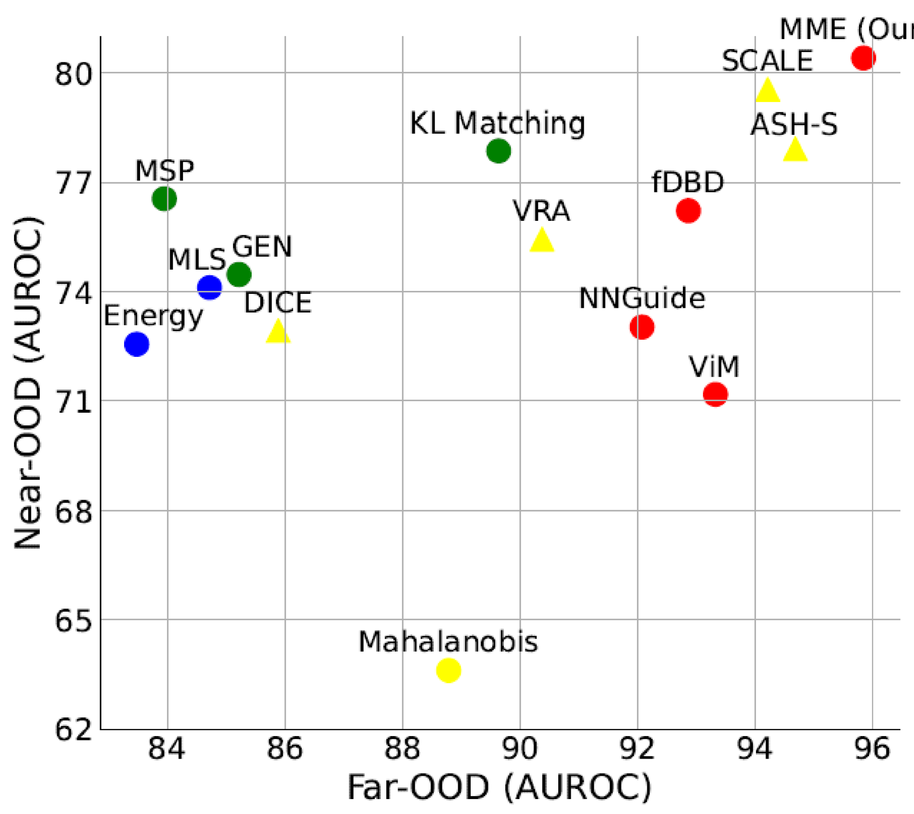
Overview of MME method



3. Extensive benchmarking results

Evaluation of various scenarios

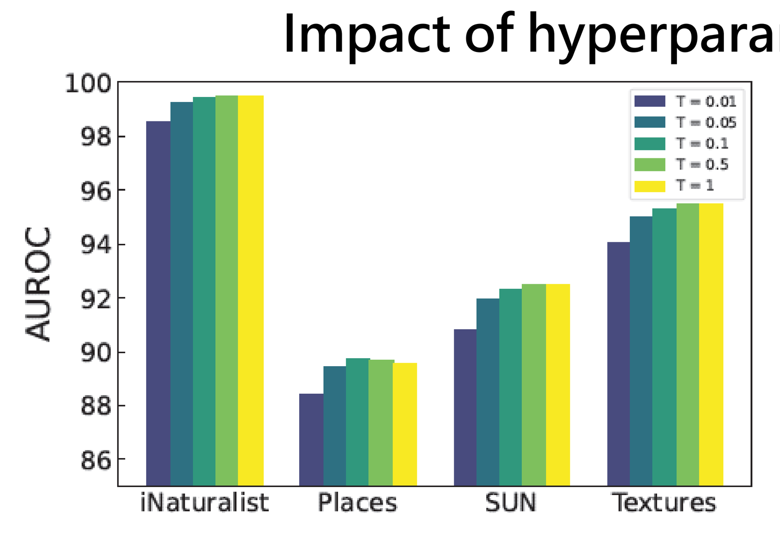


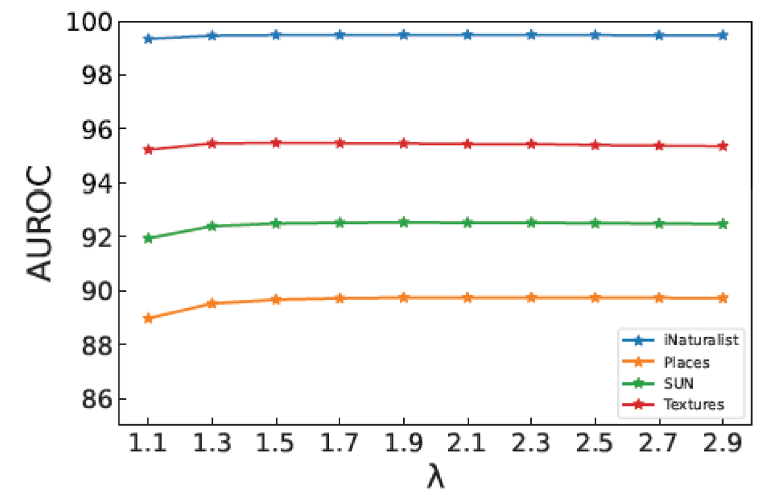


ImageNet-1K benchmark

Method	iNaturalist		Places		SUN		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MSP (ICLR'17)	87.90	64.58	79.27	79.00	81.82	71.82	79.76	77.17	82.19	73.14
MLS (ICML'22)	86.78	70.51	82.58	74.36	86.74	63.49	81.65	73.60	84.44	70.49
Energy (NeurIPS'20)	84.52	74.96	82.56	73.21	87.17	60.81	81.10	73.91	83.84	70.72
ReAct (NeurIPS'21)	91.50	48.65	87.96	54.67	92.33	38.99	90.65	50.14	90.61	48.11
Maha (NeurIPS'18)	85.80	64.69	73.06	82.35	78.10	73.38	97.33	13.95	83.57	58.59
KL (ICML'22)	90.84	<u>38.19</u>	73.17	79.10	76.43	75.88	83.22	55.16	80.91	62.08
ViM (CVPR'22)	89.38	55.09	83.33	67.94	87.85	57.79	**98.92**	**4.63**	89.87	46.36
DICE (ECCV'22)	91.44	43.64	<u>88.99</u>	<u>49.20</u>	<u>92.52</u>	35.54	83.67	55.10	89.16	45.87
ASH-S (ICLR'23)	<u>93.87</u>	<u>36.85</u>	88.49	<u>49.17</u>	91.77	<u>34.69</u>	<u>97.62</u>	<u>12.03</u>	<u>92.94</u>	<u>33.19</u>
SHE (ICLR'23)	77.65	72.25	80.31	71.63	84.65	56.68	84.58	48.84	81.80	62.35
GEN (CVPR'23)	88.67	68.32	80.19	79.66	84.17	72.23	81.48	77.91	83.63	74.53
PCA (ICCV'23)	50.37	99.77	42.19	99.80	50.15	99.25	69.69	85.35	53.1	96.04
NNGuide (ICCV'23)	88.48	52.83	88.17	50.76	**92.92**	<u>33.58</u>	94.94	21.51	91.11	40.60
SCALE (ICLR'24)	<u>93.77</u>	39.37	<u>88.96</u>	50.10	92.39	35.28	96.11	18.99	<u>92.81</u>	<u>35.94</u>
VRA (NeurIPS'24)	91.50	48.65	87.96	54.67	92.33	38.99	90.65	50.14	90.61	48.11
fDBD (ICML'24)	91.71	51.42	84.20	67.45	88.47	56.03	95.45	24.38	89.96	49.82
MME (Ours)	**95.53**	**26.91**	**89.73**	**45.03**	<u>92.70</u>	**30.64**	<u>98.45</u>	<u>7.71</u>	**94.10**	**27.57**

Impact of hyperparameters





Performance analysis

- ▷ The ensembling strategy enables MME to deliver state-of-the-art performance.
- ▷ MME consistently ranks among the top methods across models and datasets.
- ▷ Minor performance variation is observed under different hyperparameter settings.

Conclusion and perspectives

- ▷ Combining feature truncation and scoring functions enhances OOD detection performance both theoretically and empirically.
- ▷ The proposed MME strategy fuses leading OOD detectors to better address diverse OOD scenarios.
- ▷ MME achieves state-of-the-art results across multiple models, datasets, and benchmark scales.
- ▷ Thus, exploring broader OOD libraries such as OpenOOD [2] or scio [3] will further expand the design space for robust and efficient OOD detection.

[1] J. Yang, "Generalized Out-of-Distribution Detection: A Survey," in IJCV, 2024.

[2] J. Yang, "OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection," in DMLR, 2025.

[3] E. Goudout, "scio: Confidence scores for Neural Networks, made easy!," 2025. [Online]. Available: github.com/ThalesGroup/scio