




Project 9: Advanced Out-of-Distribution Detection for Multi-Class Classification

Presented by:
Mattia Raffaele Ricciardelli
Camille Justine Gomez Eugenio (2216495)

Course: Computer Vision - A.Y. 2024/2025
Engineering In Computer Science And Artificial Interlligence

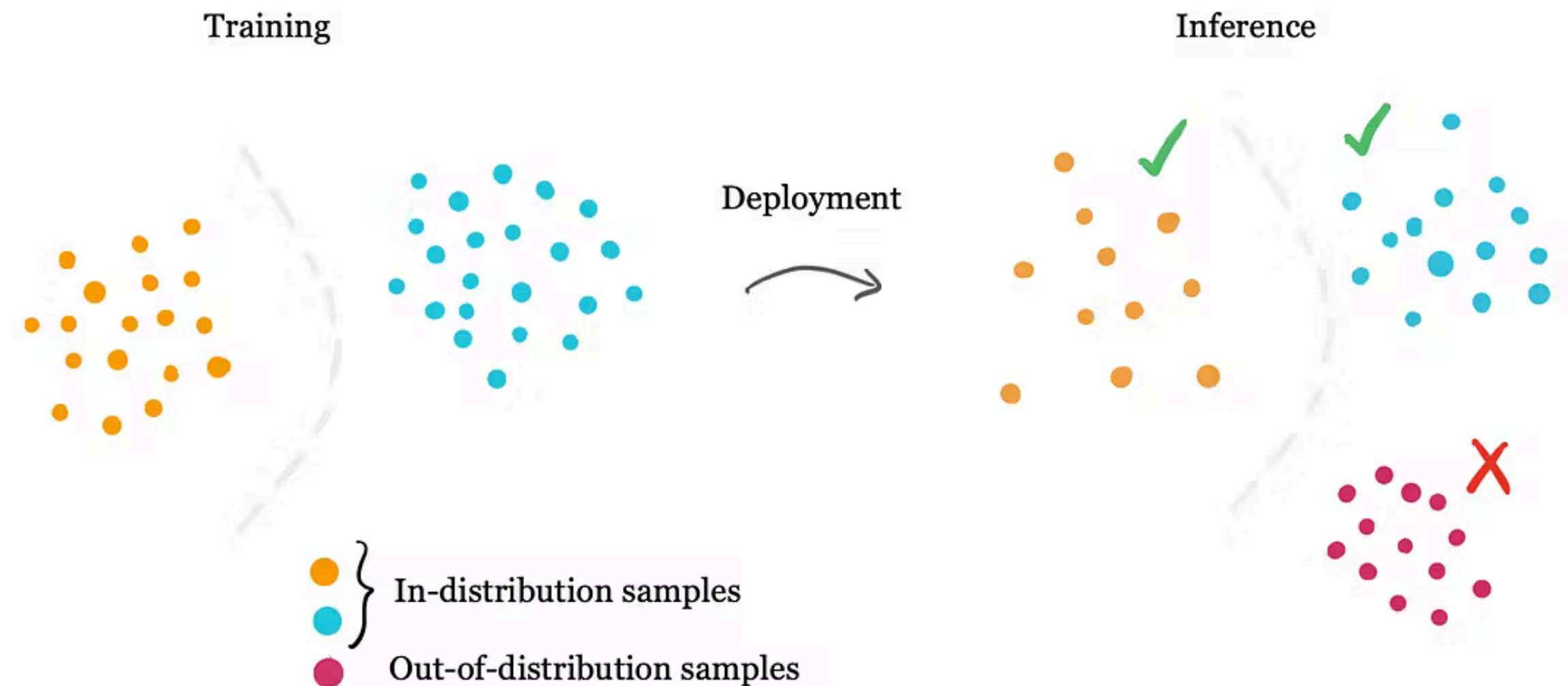


Outline

- The OOD Problem
 - OOD Detection: common approaches
 - Project Objective
 - Datasets
 - Model Architecture
 - Evaluation Metrics
 - Training Enhancements
 - Strategy 1: Baseline model (ID Only)
 - Strategy 2: Outlier Exposure
 - Results Discussion
 - Conclusions
- 

The OOD Problem

Detecting **out-of-distribution** (OOD) inputs is essential for building reliable deep learning systems. In real-world settings, models often face unfamiliar inputs that can lead to incorrect or unsafe predictions: making robust OOD detection a critical requirement.






OOD Detection: common approaches

Various methods have been proposed to identify OOD inputs by analyzing the model's output confidence, internal representations, or sensitivity to perturbations.

Some strategies:

- **Energy-Based Detection**: Computes an energy score from the model's logits to distinguish ID from OOD inputs. Lower energy indicates high confidence (ID), while higher energy suggests uncertainty (OOD). *[Liu et al., NeurIPS 2020]*
 - **Gradient-Regularized Detection**: Enhances model robustness by penalizing sharp gradients with respect to the input, encouraging smoother and more reliable predictions for ID data while increasing sensitivity to OOD inputs. *[Sharifi et al., arXiv 2024]*
 - **CORES (Convolutional Response-based Score)**: Leverages intermediate convolutional feature responses to assess distributional mismatch, producing a class-agnostic score based on activation consistency across channels. *[Tang et al., CVPR 2024]*
- 



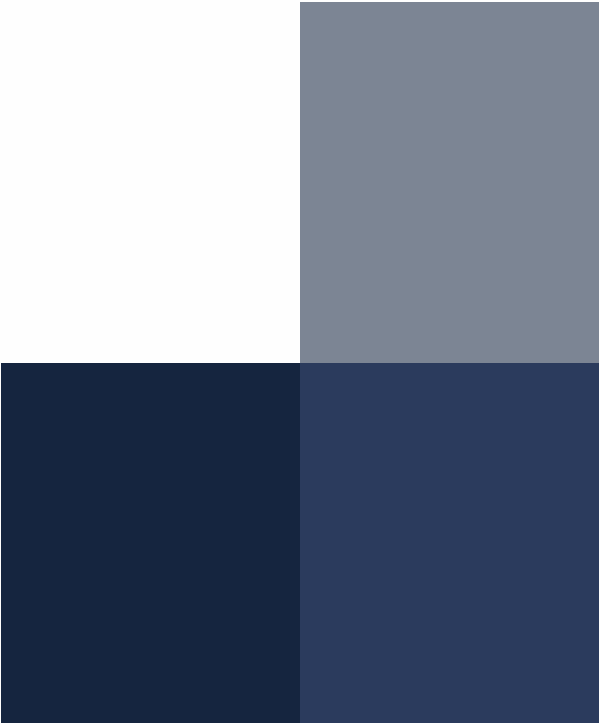
Project Objective

The goal of this project is to build a robust image classification system that can also detect out-of-distribution (OOD) inputs.

To achieve this, we:

- **Defined** a multi-class classification model using **ResNet-50**, trained on **Food-101**
- **Integrated** an OOD detection method based on energy scores, evaluated both in baseline and Outlier Exposure settings
- **Evaluated** performance using standard OOD detection metrics such as **accuracy**, **AUROC**, **AUPR**, and **FPR@95TPR**

As a final robustness test, we challenged the model on an unseen dataset (the Oxford 102 Flower dataset) to assess its generalization capabilities to completely different visual domains.



Datasets



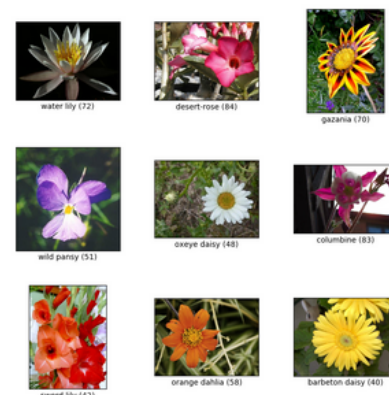
In-Distribution (ID): Food-101

- 101 food categories with high-resolution, high-variance images
- Realistic and complex visual features
- Used to train and validate the classification model



Out-of-Distribution (OOD): SVHN

- Street View House Numbers – images of digits in natural scenes
- Completely different domain from Food-101
- Used for evaluating the OOD detection capability

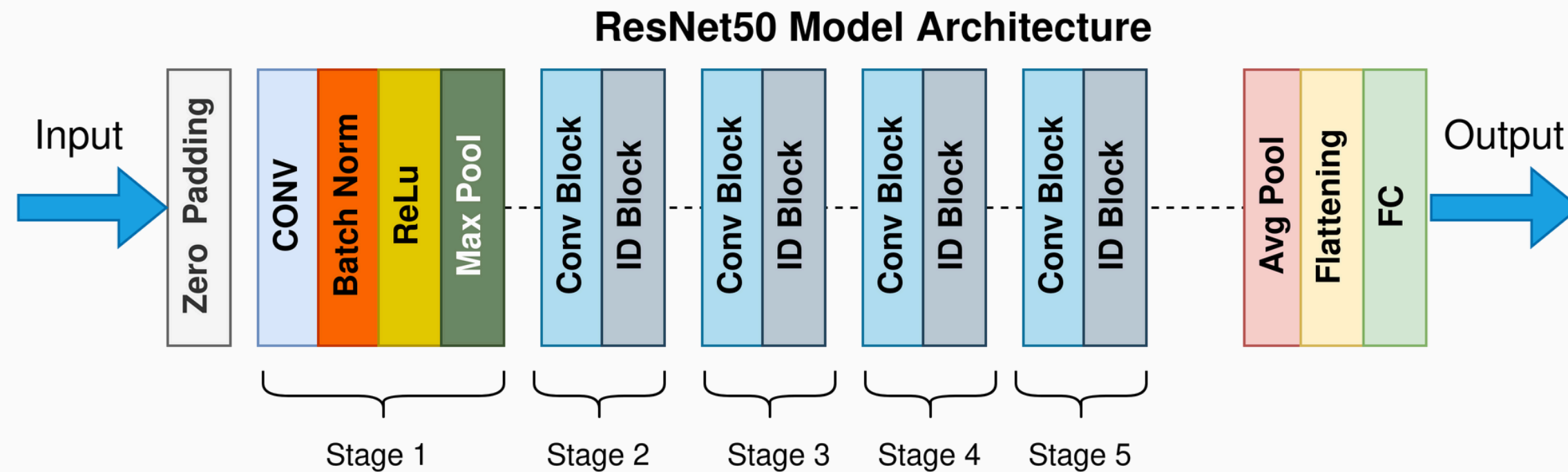


Unseen OOD (for robustness test): Oxford 102 Flowers

- 102 flower categories with fine-grained visual differences
- High intra-class similarity and distinct from both Food-101 and SVHN
- Used only to test the robustness and generalization of the final model to novel, unseen domains

Model Architecture

We used **ResNet50**, a deep convolutional neural network **pretrained** on ImageNet, as the backbone for our classifier. The model was fine-tuned on the Food-101 dataset using **cross-entropy loss**.



Evaluation Metrics

- **Accuracy**: the proportion of correct predictions over all predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **AUROC (Area Under the ROC Curve)**: measures the model's ability to distinguish between in-distribution and out-of distribution samples across all thresholds:

$$\text{AUROC} = \int_0^1 \text{TPR}(FPR) dFPR$$

- **AUPR-In**: the area under the precision-recall curve when in-distribution is considered the positive class

$$\text{AUPR-In} = \int_0^1 \text{Precision}(Recall) dRecall$$

- **FPR@95TPR**: the false positive rate when the true positive rate (recall) is fixed at 95%

$$\text{FPR@95TPR} = \left. \frac{FP}{FP + TN} \right|_{\text{TPR}=95\%}$$

Training strategy: overview

To assess the effectiveness of OOD detection, we implemented and compared two training strategies:

1

Baseline Training (ID Only)

- Train the classifier on in-distribution (Food-101) data only
- Evaluate OOD detection using energy scores without seeing OOD samples during training

2

Outlier Exposure (ID + OOD)

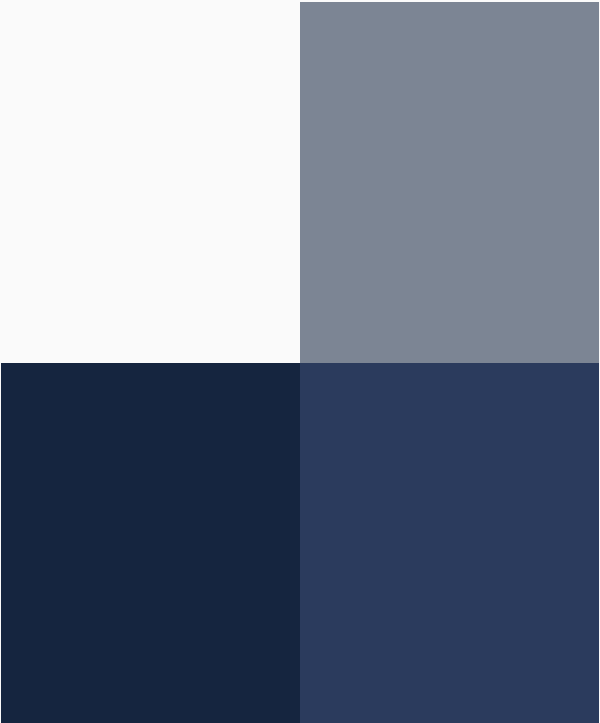
- Fine-tune the model using a mix of ID and OOD data (Food-101 + SVHN)
- Use energy-based hinge loss to explicitly separate distributions
- This two-stage approach allows us to evaluate how direct exposure to OOD samples improves detection performance.

At last, we tested the fully trained model on the unseen Oxford 102 Flowers dataset to assess its robustness and ability to generalize to entirely new, fine-grained visual categories beyond both Food-101 and SVHN.



Training Enhancements

To improve **generalization** and **calibration**, during training we applied:

- **Early Stopping** with patience to prevent overfitting
 - **Temperature Scaling** to calibrate OOD detection scores: logits are divided by a fixed temperature T before computing the energy and maximum softmax probability (MSP) scores.
 - **Weight Decay** and data augmentation for regularization
 - During Outlier Exposure, the same architecture was fine-tuned using a **cross entropy** + **combined energy-based loss**.
- 



Strategy 1: Baseline model (ID only) Training

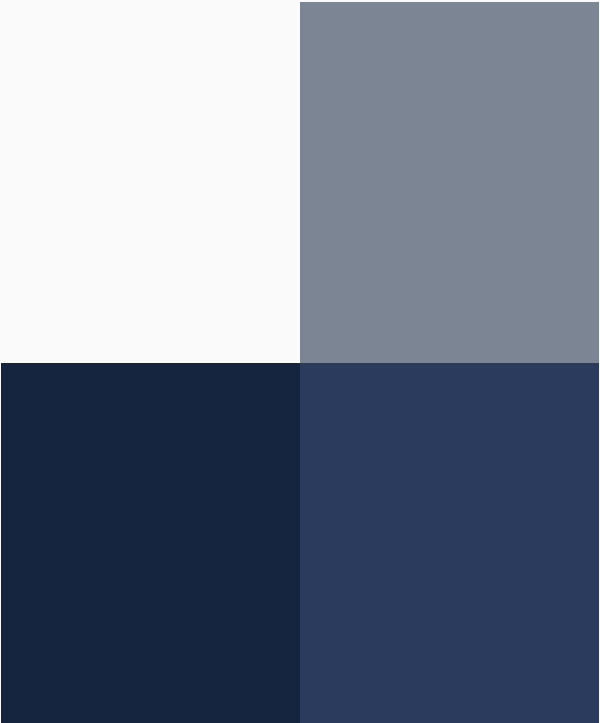
Phase 1 – Training on In-Distribution Only

We begin by training a standard ResNet-50 model on the Food-101 dataset, with no out-of-distribution data or auxiliary losses.

Objectives:

- Learn to classify 101 food categories effectively
- Establish a “vanilla” baseline to test OOD behavior later

Training Details:

- **Loss:** Cross-entropy with label smoothing ($\epsilon = 0.1$)
 - **Optimizer:** Adam
 - **Scheduler:** Halve learning rate every 5 epochs
 - **Softmax** outputs used for prediction
 - Progress tracked via mini-batch loss & top-1 accuracy
- 

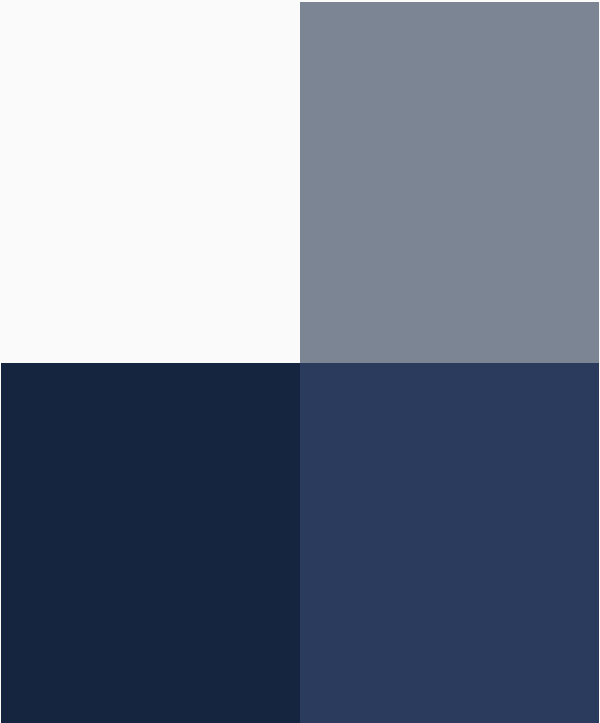


Strategy 1: Baseline model (ID only) Evaluation

Phase 1 – OOD Evaluation Without Exposure

After training on Food-101, we evaluate the model's ability to detect OOD inputs (without ever seeing them during training)

Evaluation Procedure:

- Test classification accuracy on Food-101 (ID test set)
 - Feed SVHN images as OOD inputs
 - Compute two detection scores:
 - **Energy score**
 - **Max softmax probability (MSP)**
 - Both scores are calibrated using temperature scaling
- 

Strategy 2: Outlier Exposure Training

Phase 2 – Fine-Tuning with Outlier Exposure (OE)

In this phase, we fine-tune the ResNet-50 model with mixed batches of Food-101 (ID) and SVHN (OOD) images. We introduce an energy-based auxiliary loss to push OOD samples away from the ID manifold, while still learning to classify the 101 food categories.

Total loss: $L = L_{\text{CE}} + \lambda \cdot (L_{\text{OOD}} + \alpha \cdot L_{\text{ID}})$

Energy Score: $E(z) = -T \cdot \log \sum_c \exp(z_c/T)$

High penalties:

- $L_{\text{ID}} = \max(E_{\text{ID}} - m_{\text{in}}, 0)^2$
- $L_{\text{OOD}} = \max(m_{\text{out}} - E_{\text{OOD}}, 0)^2$

Training Enhancements:

- **λ ramp-up** during warm-up and early epochs
- **Label smoothing decay** after ramp phase
- **CosineAnnealingWarmRestarts** scheduler ($T_0 = 5$, $T_{\text{mul}} = 2$)
- **Early stopping** based on validation loss
- **BatchNorm recalibration** post-training to improve test-time stability



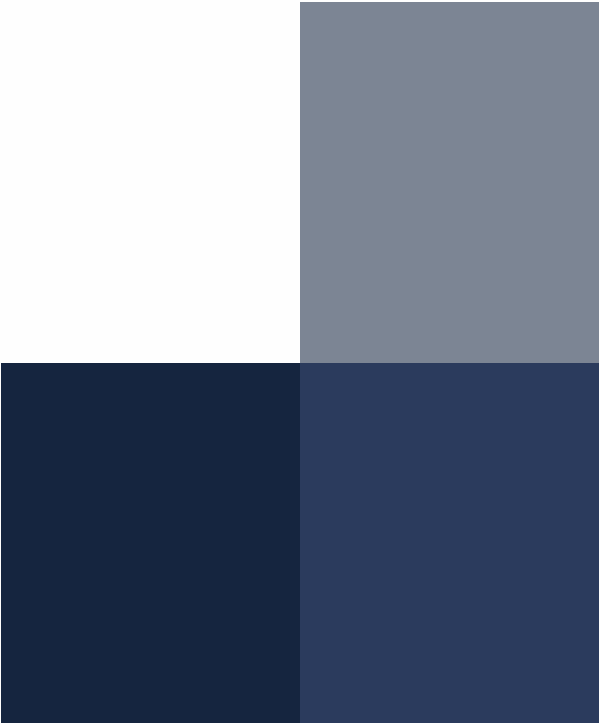
Strategy 2: Outlier Exposure Evaluation

Phase 2 – Evaluation of OOD-Aware Classifier

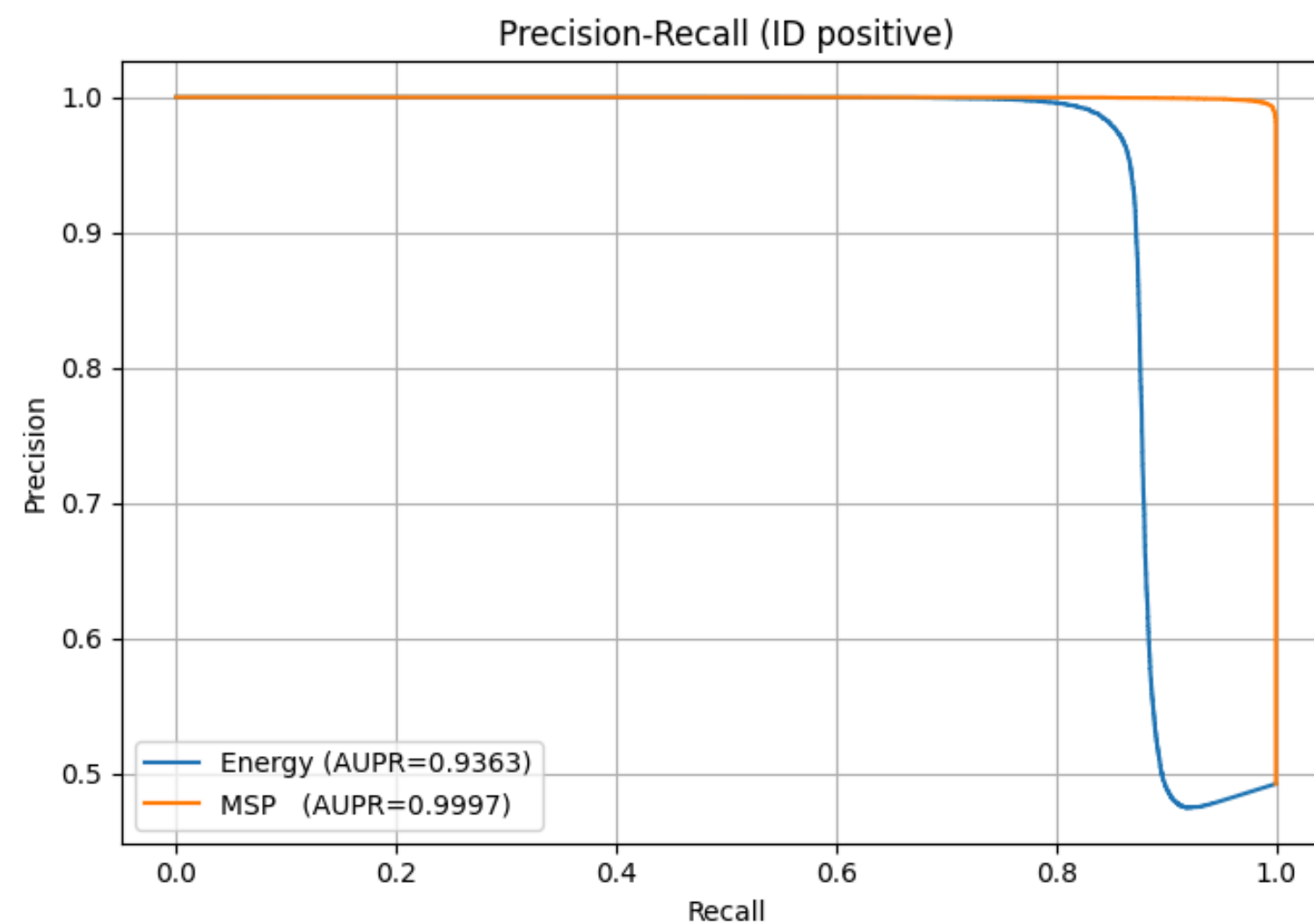
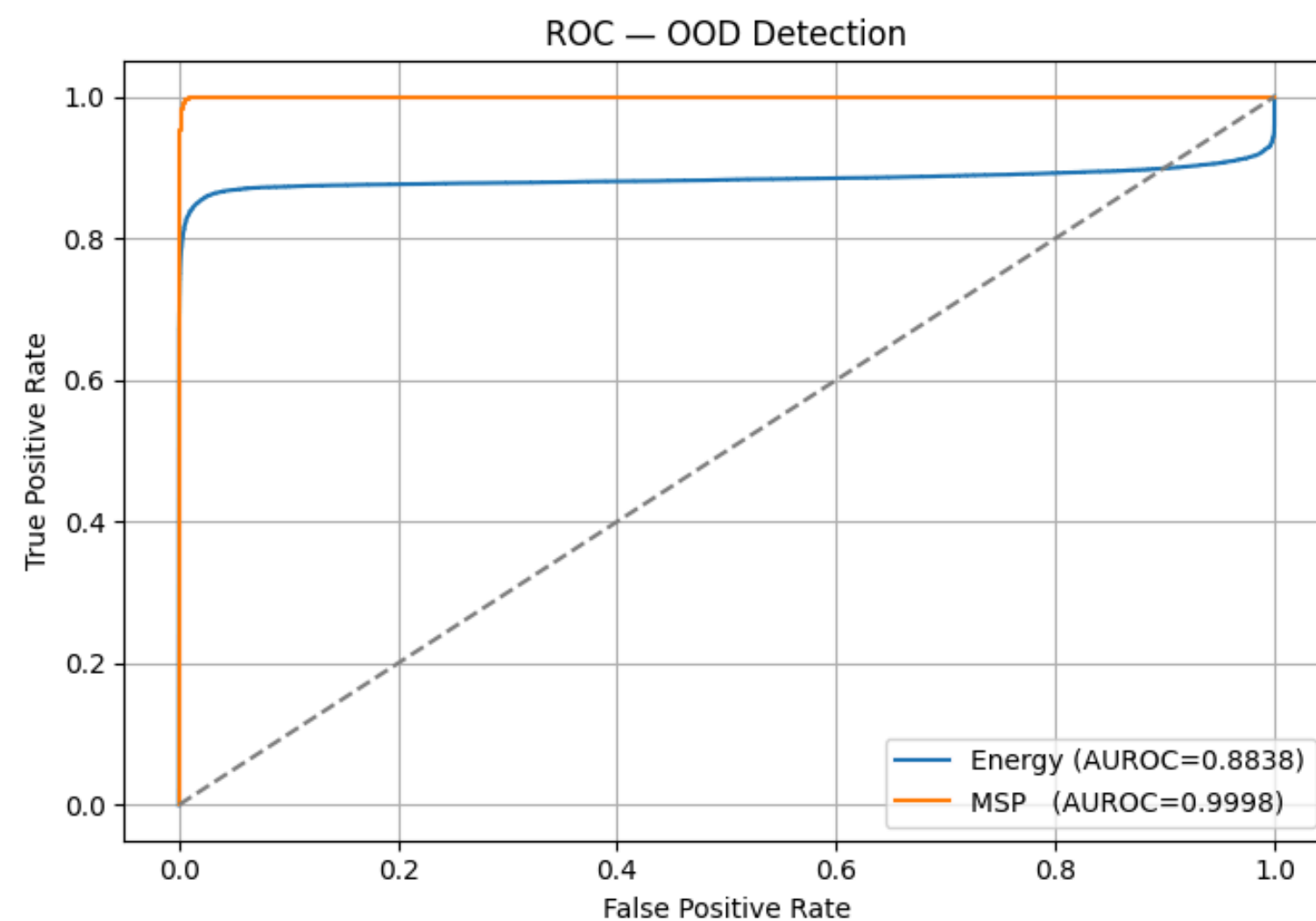
We evaluate the fine-tuned model on both in-distribution (Food-101) and out-of-distribution (SVHN) data.

The goal is to measure whether the model has learned to assign high energy to OOD inputs.

Evaluation Steps:

- Compute accuracy on Food-101 test set (ID)
 - Collect **energy** and **MSP** scores on both ID and OOD data
 - Use **temperature scaling** to normalize logits before scoring
- 

Results: Baseline Model (ID only)

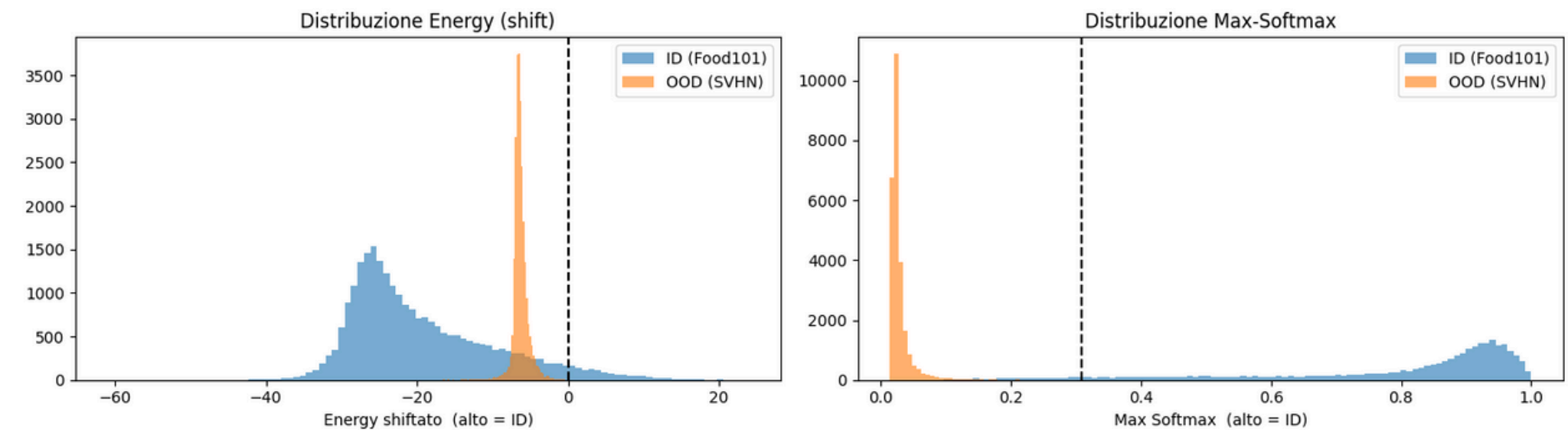


Results: Baseline Model (ID only)

METRICHE OOD (ID = Food-101, OOD = SVHN)

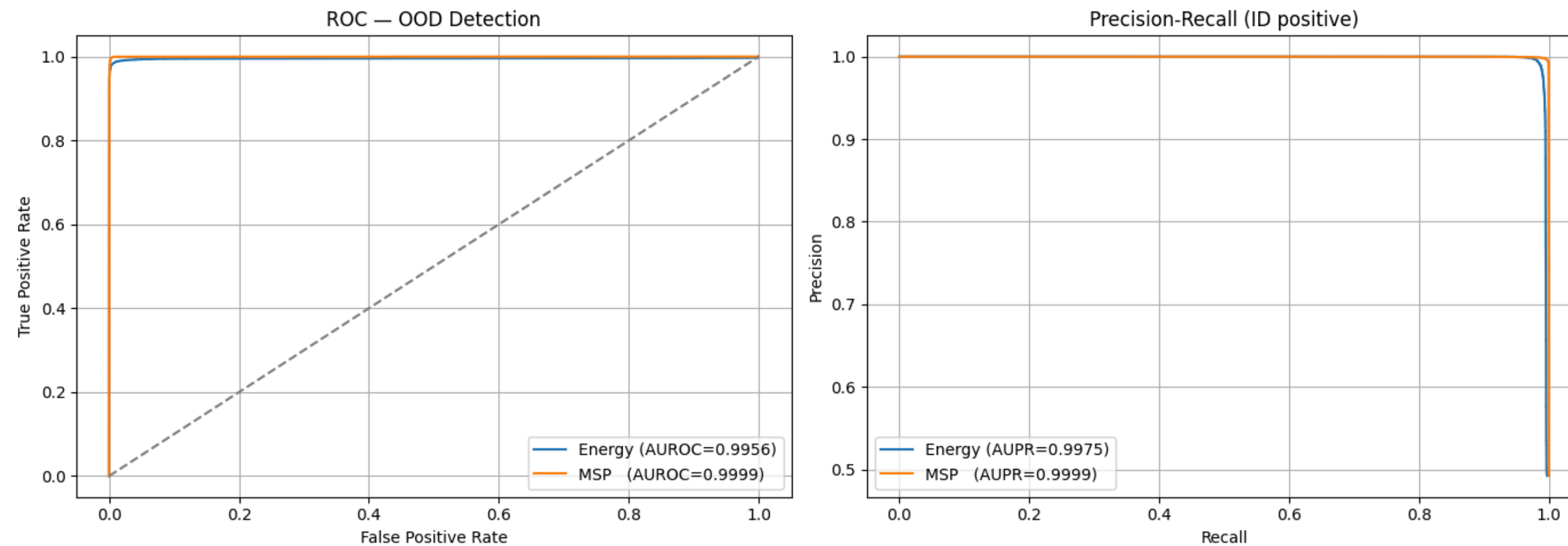
AUROC (Energy)	: 0.8838
AUROC (Soft-max)	: 0.9998
FPR@95TPR (Energy)	: 0.03%
FPR@95TPR (Soft-max)	: 0.13%
AUPR-In (Energy)	: 0.9363
AUPR-In (Soft-max)	: 0.9997

final test accuracy: 86.78%



- Although the softmax-based OOD detection seems to perform very well on this test set, the energy score provides more stable and realistic OOD detection.
- These baseline results highlight the limitations of training a classifier only on ID data. This sets the stage for the Outlier Exposure phase, which improves both classification and OOD metrics.

Results: Outlier Exposure

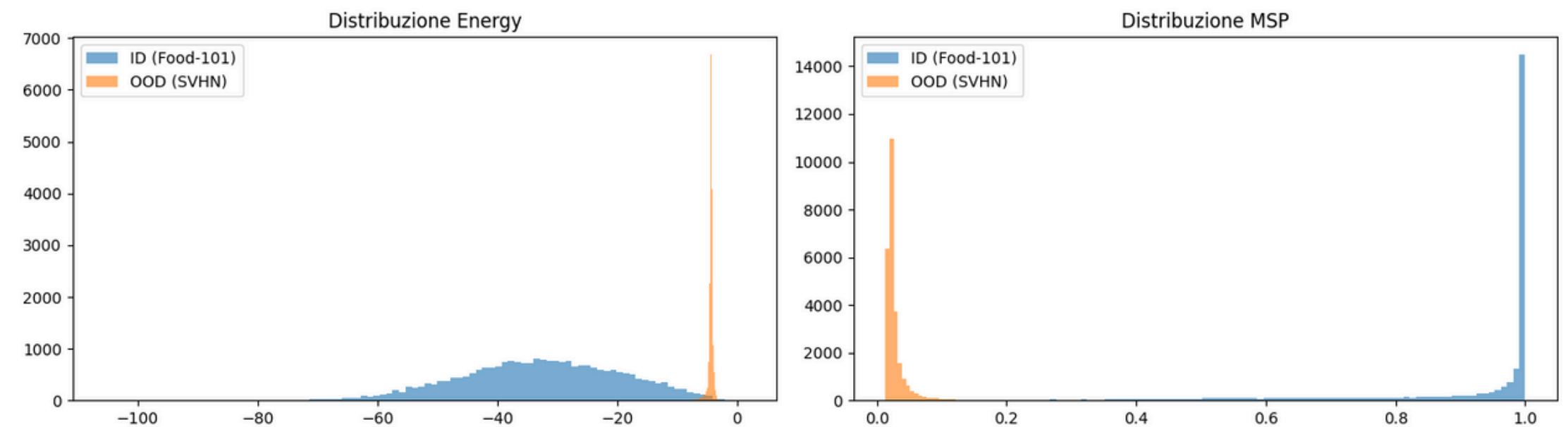


Results: Outlier Exposure

METRICHE OOD

AUROC (Energy)	: 0.9956
AUROC (Soft-max)	: 0.9999
FPR@95TPR (Energy)	: 0.02%
FPR@95TPR (Soft-max)	: 0.04%
AUPR-In (Energy)	: 0.9975
AUPR-In (Soft-max)	: 0.9999

final test accuracy: 87.44%



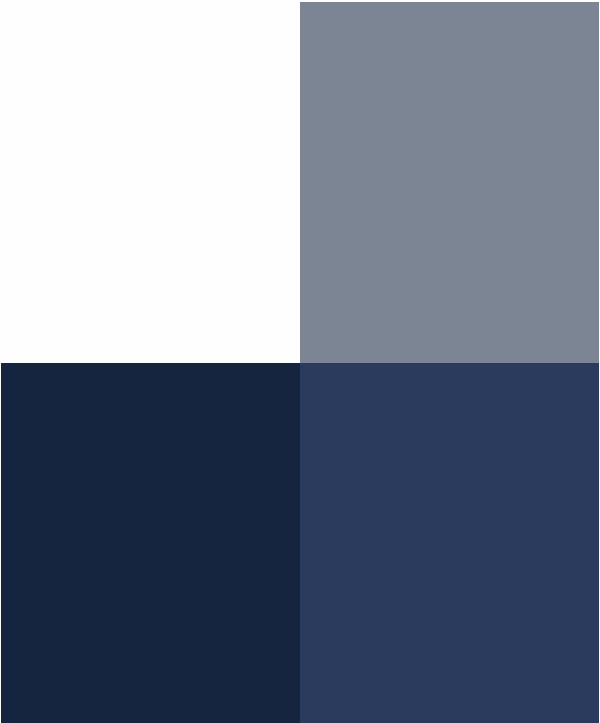
Outlier Exposure significantly enhances the model's OOD detection capabilities while slightly improving classification performance. This validates the effectiveness of fine-tuning with OOD data using an energy-based loss.



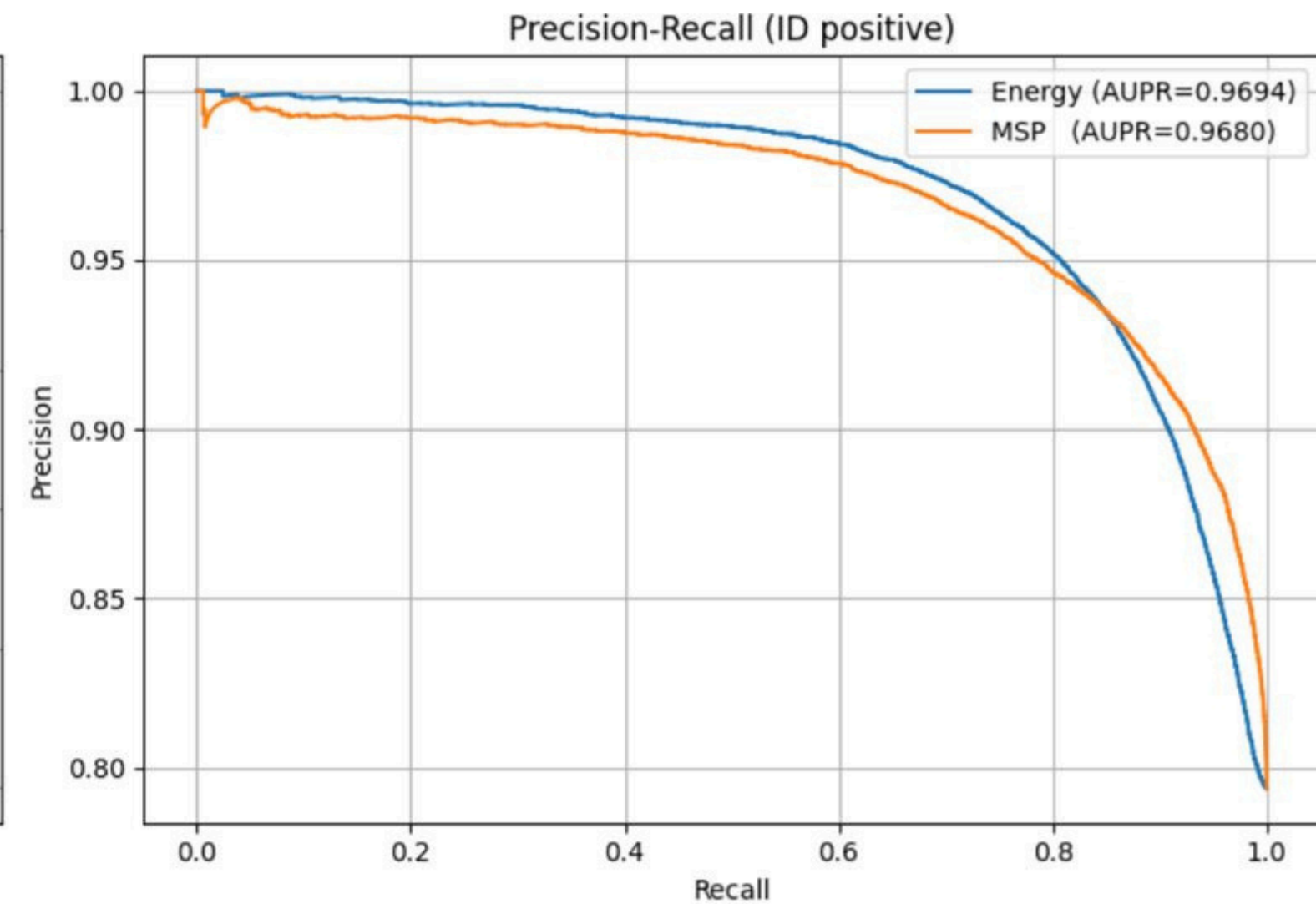
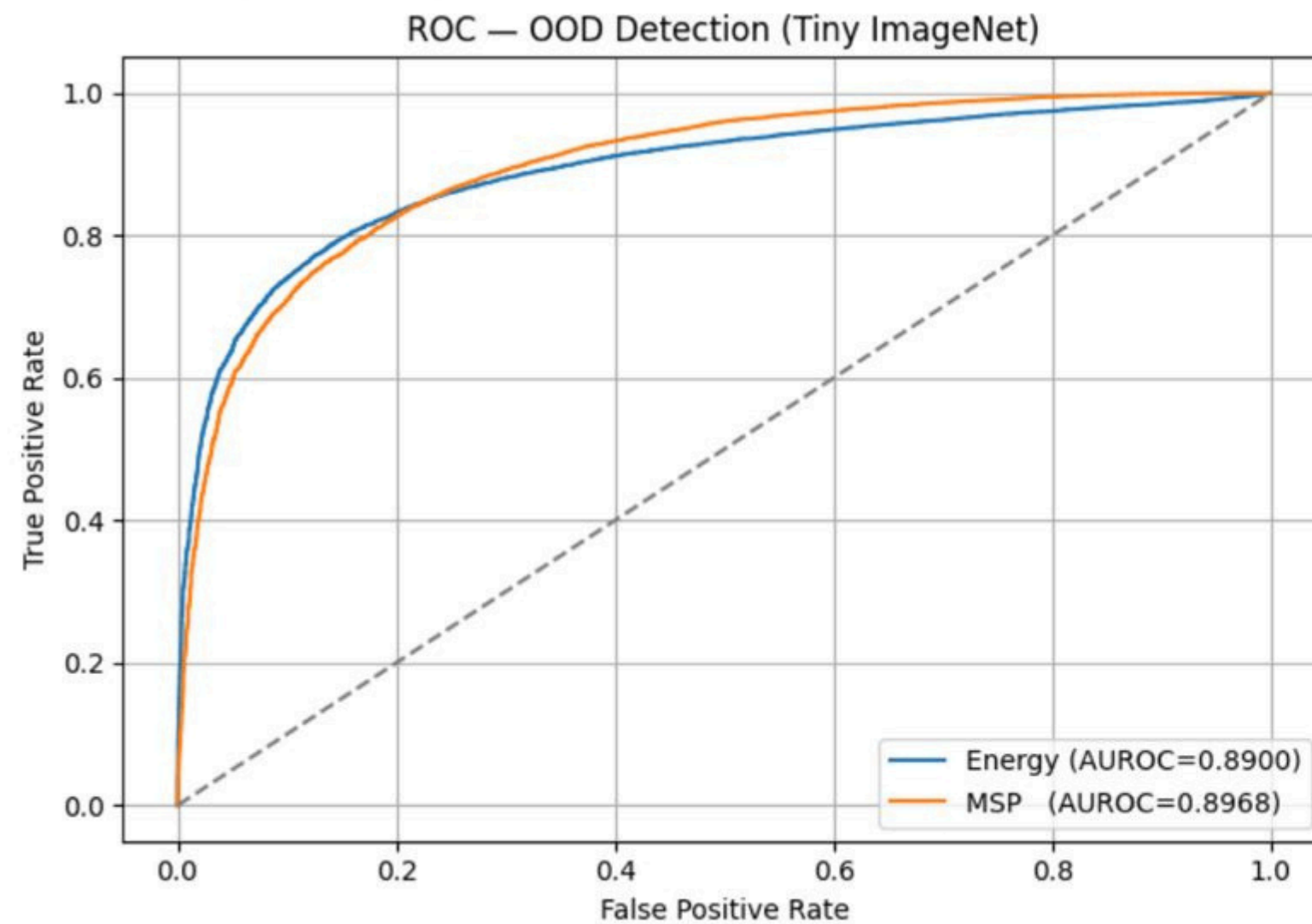
Results: Comparison

Metric	Baseline	Outlier Exposure
AUROC (Energy)	0.8838	0.9956
AUROC (Softmax)	0.9998	0.9999
FPR@95TPR (Energy)	0.03%	0.02%
FPR@95TPR (Softmax)	0.13%	0.04%
AUPR-In (Energy)	0.9363	0.9975
AUPR-In (Softmax)	0.9997	0.9999

Outlier Exposure improves OOD detection across all metrics while also slightly increasing in-distribution classification accuracy.

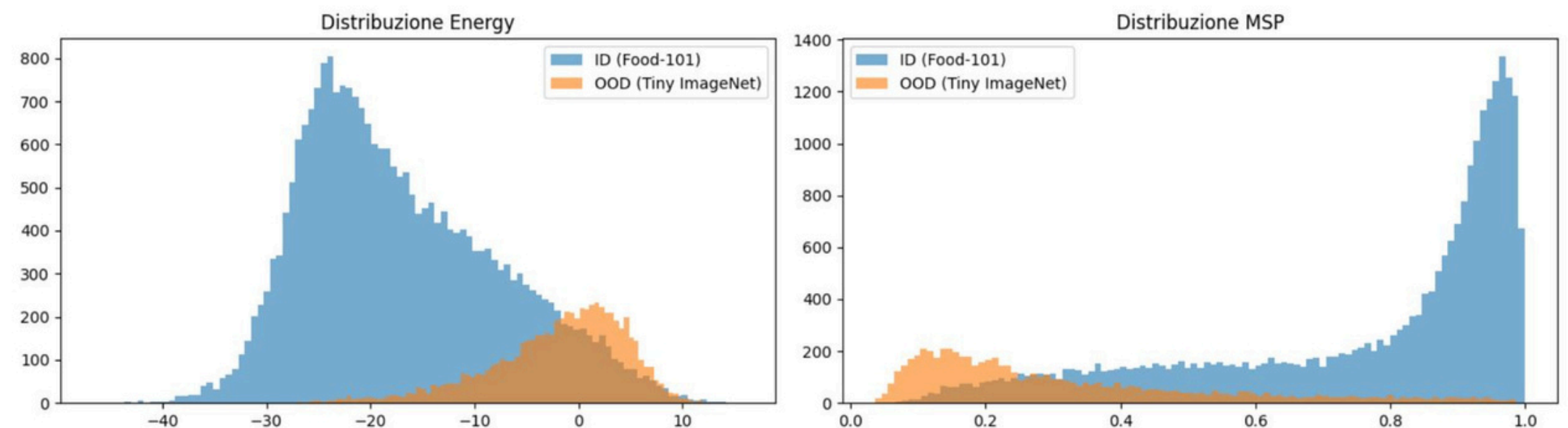


Results: Evaluation On Unseen Dataset



Results: Unseen Dataset

Metric	Energy Score	Softmax Score
AUROC	0.89	0.8968
FPR@95TPR	60.79%	47%
AUPR-In	0.9694	0.968



Despite a drop in some OOD metrics (notably FPR@95TPR), the model maintains reliable detection capability on completely novel data, confirming the effectiveness and robustness of energy-based OOD detection even under distribution shift.

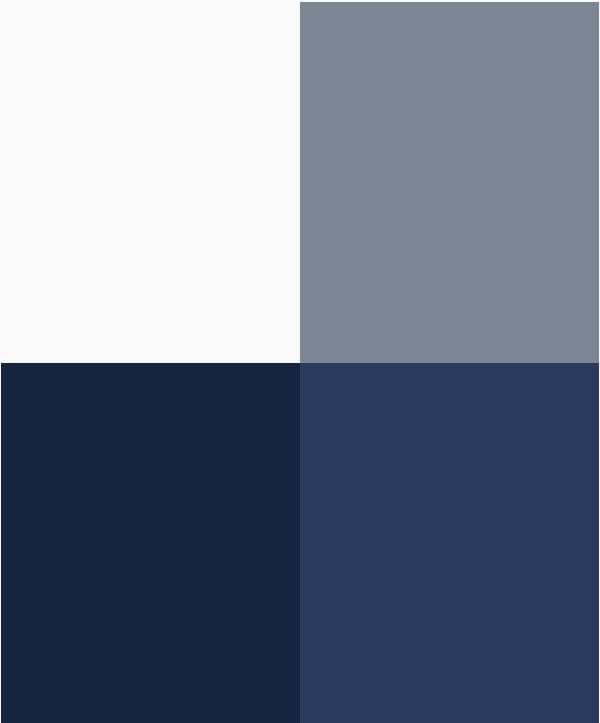




Conclusion

This project validated the effectiveness of energy-based OOD detection on a ResNet50 trained on Food-101, with SVHN and Oxford Flowers as OOD datasets.

The model showed strong performance across AUROC, AUPR, and FPR@95TPR, but there's room for further improvement.

Future directions include:


- Upgrading the model architecture (e.g., EfficientNet, ViT)
 - Enhancing uncertainty calibration (e.g., multi-temperature scaling)
 - Using domain-aware or synthetic OOD augmentations
 - Adopting advanced OOD methods (e.g., ODIN, Mahalanobis, GradNorm)
 - Applying self-supervised pretraining for better generalization
- 



Thank you
for your
attention!



Resources

- Liu, W., Wang, X., Owens, J. D., & Li, Y. (2020). Energy-based Out-of-distribution Detection.
 - Sharifi, S. et al. (2024). Gradient-Regularized Out-of-Distribution Detection.
 - Tang, K., et al. (2024, June). CORES: Convolutional Response-based Score for Out-of-distribution Detection.
- 
- 