

Computer Vision 2024/2025

Advanced Out-of-Distribution Detection for Multi-Class Classification

Federico Tranzocchi, Gianmarco Corsi

≡ Outline

- **Problem statement:** Understanding the challenge of Out-of-Distribution (OOD) data.
- **State of the art:** A brief overview of current approaches to OOD detection.
- **Proposed method:** Our methodology for training, fine-tuning, and scoring.
- **Datasets & Setup:** The data and environment used for our experiments.
- **Model evaluation:** A deep dive into our results, comparing baseline vs. fine-tuned models.
- **Comparative analysis:** Energy score vs. CORES score.
- **Conclusions & Future work** Final thoughts and potential next steps.

ⓘ The OOD challenge

What happens when a model sees an “unknown” input?

Training



Model learns to classify **food**.

Inference



Model is shown a **digit**.

The problem

“It’s a hot dog!”
(99% confident)

It makes a **highly confident**, but
completely wrong prediction.

Current approaches

Post-hoc methods: analyzing a pre-trained model



Softmax score

The classic baseline.
Often overconfident.



Energy score

More robust score
derived from
model logits.



CORES score

Based on
convolutional kernel
response patterns.

These methods work without needing access to OOD data during training.

Current approaches

Fine-tuning methods: using auxiliary OOD data



Outlier exposure

Trains model to be "unconfident" on OOD samples.



Energy fine-tuning

Explicitly shapes an "energy surface" to separate ID and OOD scores.

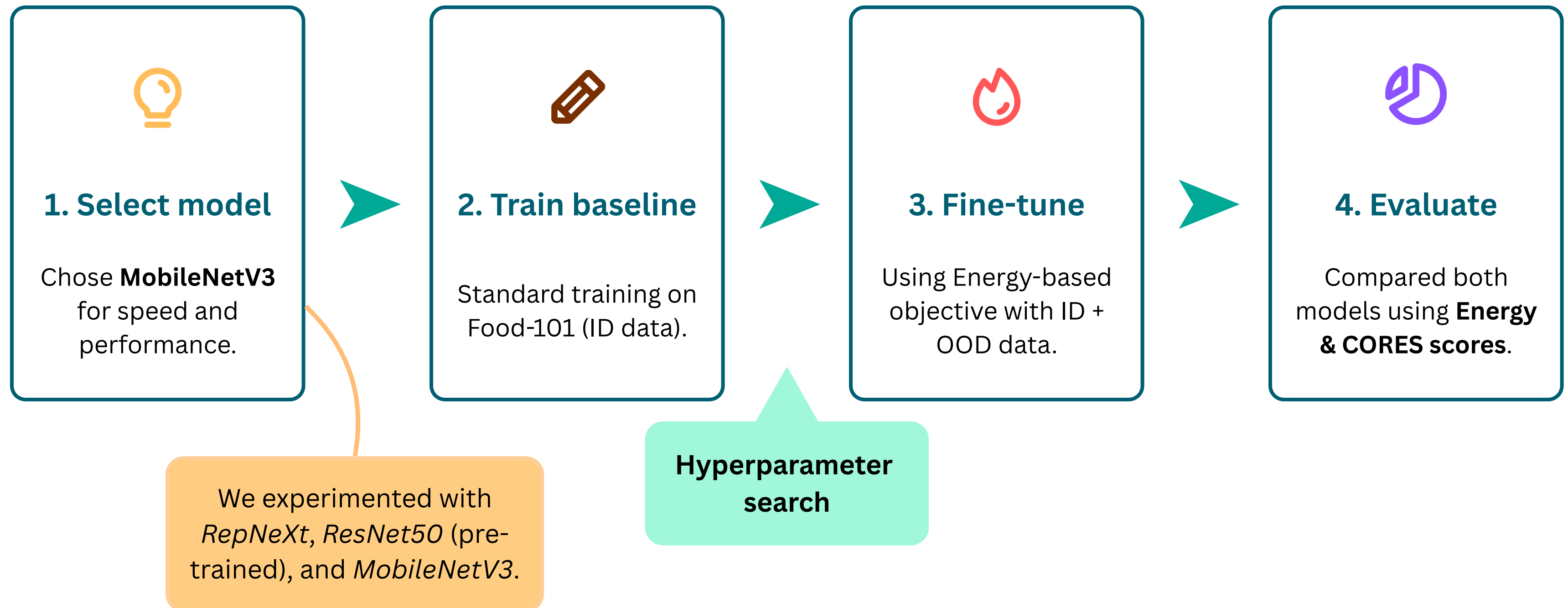


Gradient regularization

Smooths the decision boundary for better robustness.

These methods leverage a dataset of known outliers to improve the model.



↪ Our project workflow






Why MobileNetV3?

CHOSEN MODEL



RepNeXt

-  Slow training speed
-  Poor OOD performance

MobileNetV3

-  Fast to train and fine-tune
-  Excellent ID accuracy
-  Great OOD results

ResNet50

-  High score variance
-  Potential pre-train bias

MobileNetV3 architecture

Key building blocks

- ➡ **Stem:** A standard convolutional layer performing initial feature extraction.
- ↻ **Inverted residual blocks:** Expand channel dimensions, apply efficient depthwise convolution, and project back down, capturing rich features with fewer parameters.
- ⚖️ **Squeeze-and-Excite (SE):** A lightweight attention mechanism allowing the model to re-weight importance of each feature channel, focusing on what's most relevant.
- 📈 **SiLU:** A smooth activation that performs better than standard ReLU.

Why it works for us

- ✍️ **Lightweight:** Depthwise convolutions drastically reduce the number of parameters and computations.
- ⚡ **Efficient:** Designed for high performance on resource-constrained devices, translating to faster training and inference on our hardware.

Datasets and experimental setup

Datasets

- **In-Distribution (ID):**
 - **Food-101:** A challenging dataset with 101 food categories and high visual similarity between classes.
- **Out-of-Distribution (OOD):** We used a diverse set to test robustness.
 - **Far-OOD (dissimilar):** SVHN, CIFAR-10, FashionMNIST, EuroSAT.
 - **Near-OOD (similar):** Flowers102, DTD (Textures), FGVCAircraft, OxfordIIITPet.

Experimental setup

- **Model:** MobileNetV3.
- **Framework:** PyTorch.
- **Hardware:** NVIDIA RTX A6000.
- **Evaluation metrics:** AUROC, AUPR, FPR@95TPR.

OOD scoring functions

Energy score

$$E(x) = -T \log \sum \exp(f_i(x)/T)$$

- **Low energy** = In-Distribution (high confidence)
- **High energy** = Out-of-Distribution (low confidence)

More robust than softmax as it avoids the "shifting" that makes OOD samples appear overconfident.

CORES score

Leverages the observation that ID samples elicit stronger responses from a CNN's kernel.

- Analyzes the **response magnitude** (how high/low are the activation peaks).
- Considers also the **response frequency** (how often do kernels activate strongly).
- By **backtracking** from predictions, it identifies most sample-relevant kernels to compute robust OOD scores.

Q Hyperparameter search

To maximize the effectiveness of energy-based fine-tuning, we used *Optuna* to automatically search for the energy boundaries and learning rate.

Search space

- **Learning rate:** From $1e-5$ to $1e-3$.
- **In-Distribution margin:** The target upper energy boundary for ID samples. We searched from -27 to -8.
- **Out-of-Distribution margin:** The target lower energy boundary for OOD samples. We searched from -7 to -1.

Best parameters found

- **Learning rate:** $8e-4$
- **m_in:** -11
- **m_out:** -5

Evaluation: ID classification accuracy

Performance on Food-101 test set

First, we evaluate how well the models perform their primary task: classifying in-distribution images.

Baseline model

67.56%

Accuracy after standard
cross-entropy training.

Energy fine-tuned model

76.00%

Accuracy after energy-
bounded fine-tuning.

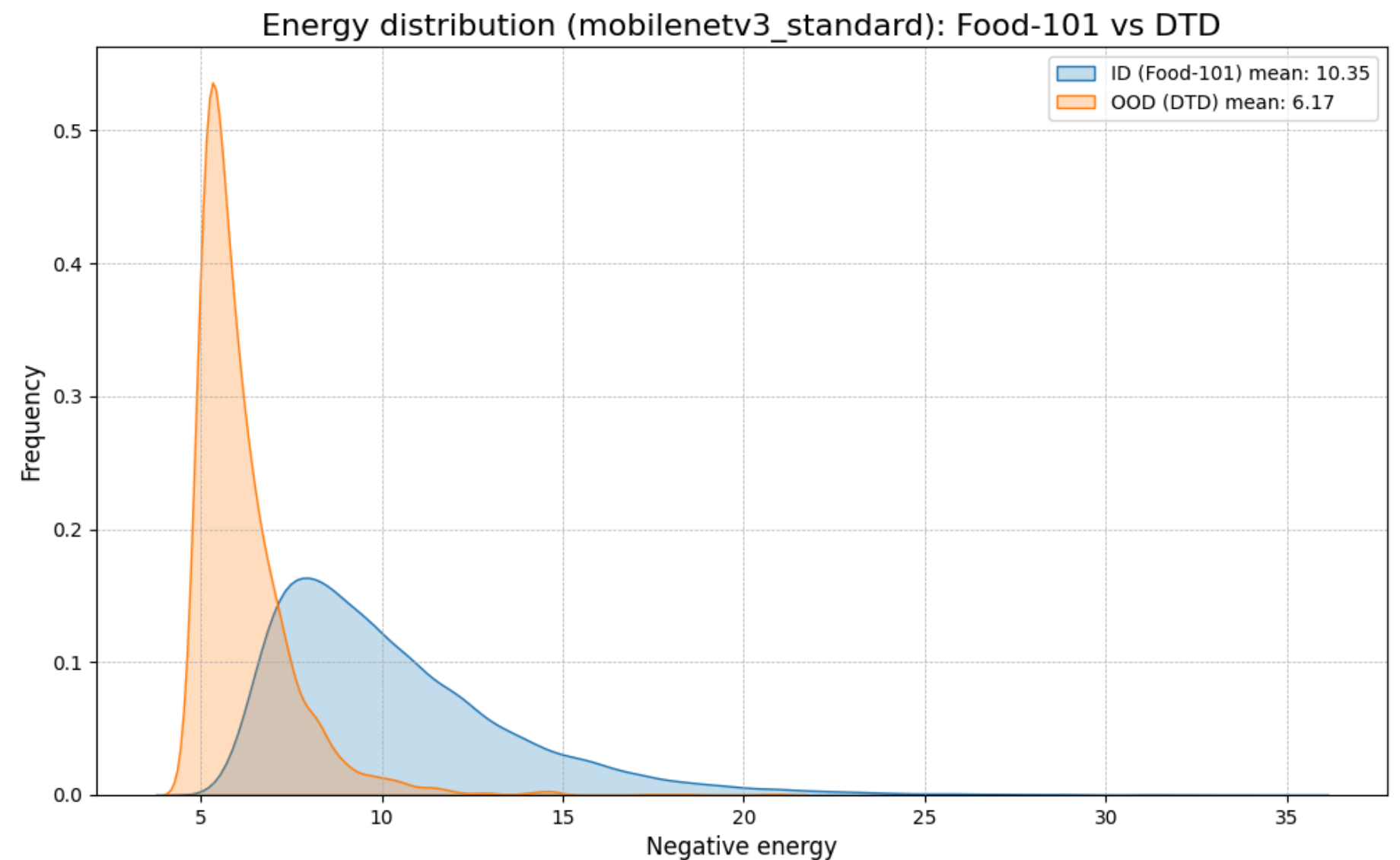
Energy-based fine-tuning doesn't just improve OOD detection; it significantly boosts the model's core classification accuracy on the ID task by **+8.44%**.

📈 Evaluation: baseline model OOD performance

OOD detection with energy score

OOD Dataset	AUROC ↑	FPR@95TPR ↓
SVHN (easy)	99.92%	0.20%
DTD (medium)	93.53%	25.32%
Flowers102 (hard)	88.25%	50.97%
OxfordIIITPet (hard)	92.39%	40.64%

The distributions for ID (blue) and OOD (orange) show some overlap, indicating confusion.

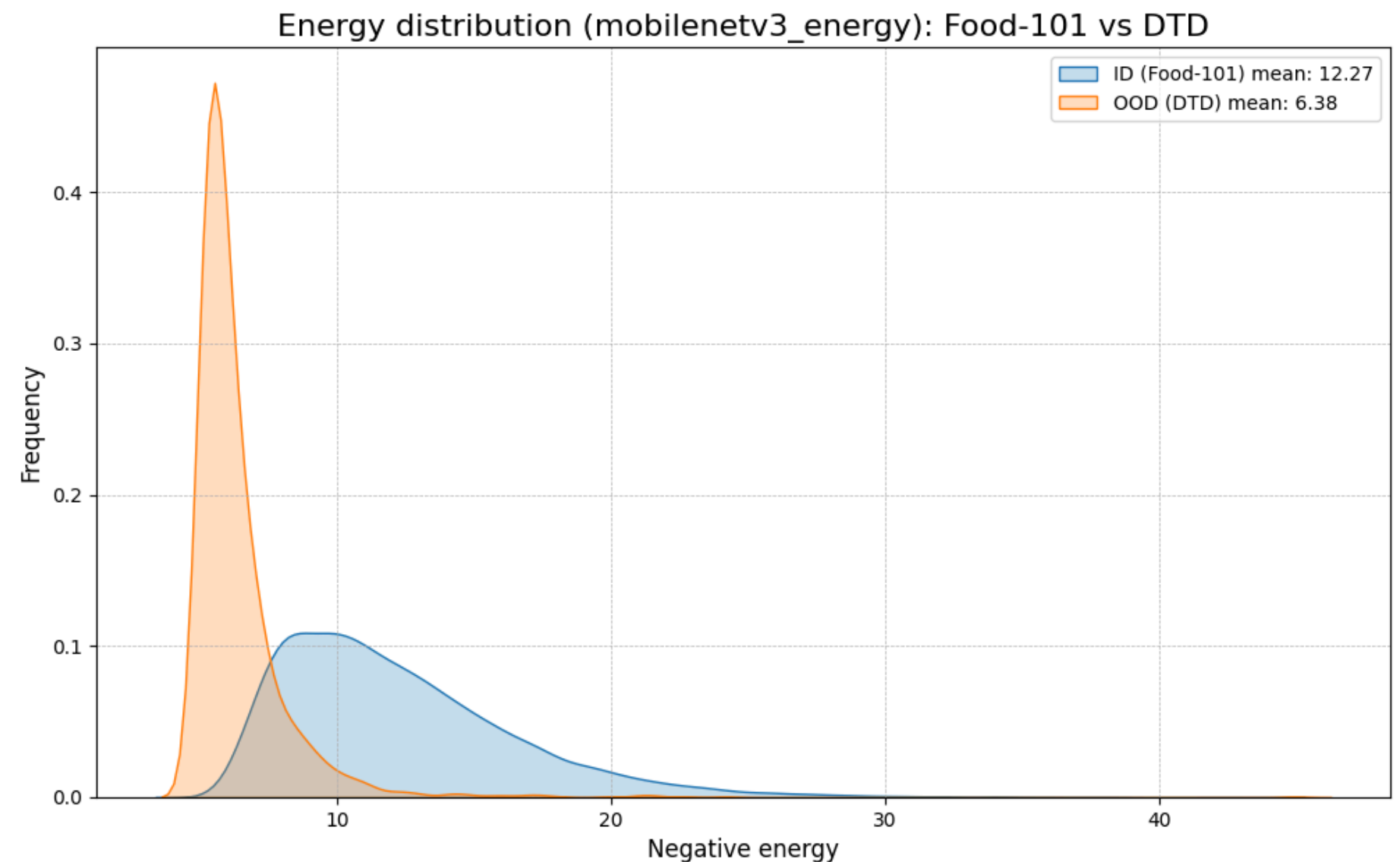


📈 Evaluation: fine-tuned model OOD performance

OOD detection with energy score

OOD Dataset	AUROC ↑	FPR@95TPR ↓
SVHN (easy)	99.96%	0.16%
DTD (medium)	95.11%	18.40%
Flowers102 (hard)	92.56%	35.92%
OxfordIIITPet (hard)	94.19%	34.29%

After fine-tuning, the model's ability to separate ID and OOD samples improves across all datasets.





Comparative analysis: baseline vs. fine-tuned

Direct comparison of FPR@95TPR (energy score)

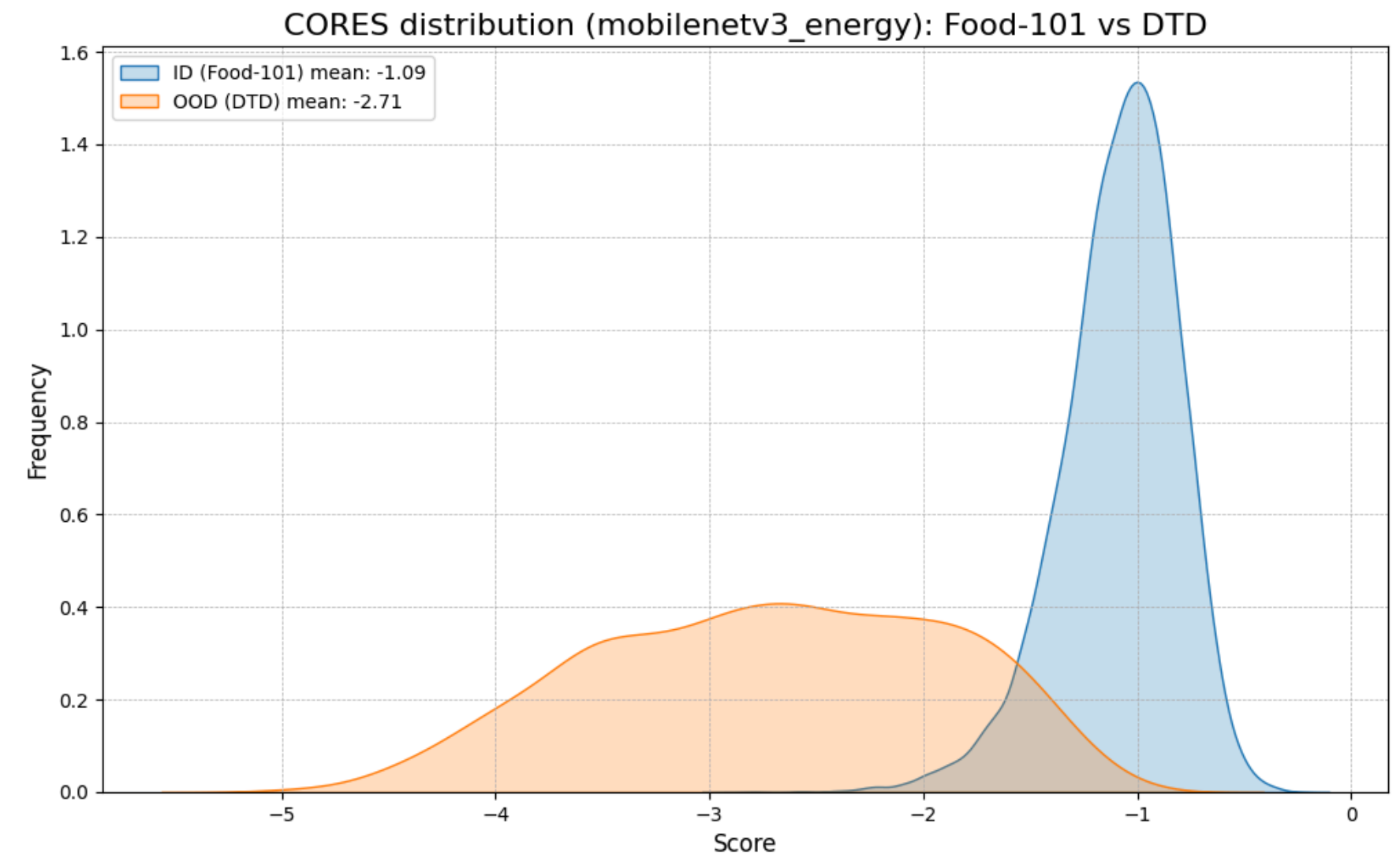
OOD Dataset	Baseline FPR@95TPR ↓	FPR@95TPR ↓	Improvement
DTD	25.32%	18.40%	-6.92%
Flowers 102	50.97%	35.92%	-15.05%
OxfordIIITPet	40.64%	34.29%	-6.35%
FGVCAircraft	1.77%	1.56%	-0.21%

- Fine-tuning provides a consistent and significant reduction in the false positive rate, especially for challenging datasets.
- The energy-bounded objective successfully pushes the ID and OOD score distributions apart, leading to a more reliable and robust OOD detector.

The largest gains are seen on the most difficult datasets.

Analysis of CORES score

Model / Score	FPR@95TPR on DTD ↓	FPR@95TPR on Flowers102 ↓
Baseline + Energy Score	25.32%	50.97%
Baseline + CORES Score	7.39%	49.93%
Fine-Tuned + Energy Score	18.40%	35.92%
Fine-Tuned + CORES Score	8.03%	43.29%



We applied the CORES scoring function to both models. It performs strongly, sometimes outperforming the Energy score on the baseline model, but shows less relative improvement after fine-tuning.

CORES vs. Energy

OOD dataset	Model / Score	AUROC ↑	FPR@95TPR ↓
DTD (Textures)	Baseline + Energy Score	93.53%	25.32%
	Baseline + CORES Score	98.51%	7.39%
	Fine-Tuned + Energy Score	95.11%	18.40%
	Fine-Tuned + CORES Score	98.36%	8.03%
Flowers 102	Baseline + Energy Score	88.25%	50.97%
	Baseline + CORES Score	89.09%	49.93%
	Fine-Tuned + Energy Score	92.56%	35.92%
	Fine-Tuned + CORES Score	92.04%	43.29%
OxfordIIITPet	Baseline + Energy Score	92.39%	40.64%
	Baseline + CORES Score	81.55%	67.59%
	Fine-Tuned + Energy Score	94.19%	34.29%
	Fine-Tuned + CORES Score	80.52%	72.96%

- **CORES** excels out-of-the-box on texture-based datasets like DTD.
- **Energy-based fine-tuning** is crucial for improving performance on visually similar datasets like **Flowers102** and **OxfordIIITPet**, where CORES struggles.

👉 Conclusions

- **MobileNetV3 is an effective backbone** for this task, providing a great trade-off between speed and performance for both classification and OOD detection.
- **Energy-based fine-tuning is highly effective.** It not only improved OOD detection metrics across a diverse set of datasets but also boosted the base model's classification accuracy by over 8%.
- The method shows its strength particularly on **near-distribution (hard) OOD datasets**, where baseline methods often struggle.
- **CORES is a powerful alternative scoring function** that works very well "out-of-the-box" without fine-tuning, especially on datasets with strong textural differences.
- A systematic approach combining a solid baseline, targeted fine-tuning, and automated hyperparameter search is key to building robust and reliable models.

Future Work

- **Explore Gradient Regularization (GReg):** Implement the GReg loss term from *Sharifi et al. (2024)* in addition to the energy loss to see if it can further improve robustness and smooth the decision boundary.
- **Combine Methods:** Investigate if fine-tuning with an objective that incorporates both Energy and CORES score properties could lead to a model that excels with both scoring functions.
- **More Complex Architectures:** Apply this methodology to larger and more complex models like Vision Transformers (ViTs) to see if the performance gains translate to different architectural paradigms.
- **Broader OOD Scenarios:** Test the final model in even more challenging scenarios, such as detecting adversarial attacks or handling data from different modalities.

References

- Liu, W., et al. (2020). **Energy-based Out-of-distribution Detection.**
- Tang, K., et al. (2024). **CORES: Convolutional Response-based Score for OOD Detection.**
- Sharifi, S., et al. (2024). **Gradient-Regularized Out-of-Distribution Detection.**
- Ramachandran et al. (2017). **Searching for Activation Functions.**
- Zhao et al. (2024). **RepNeXt: A Fast Multi-Scale CNN using Structural Reparameterization.**

Thank you!