Out-of-Distribution Detection with Adversarial Outlier Exposure

Thomas Botschen[†], Konstantin Kirchheim[†], Frank Ortmeier
firstname.lastname@ovgu.de
University of Magdeburg, Department of Computer Science

CVPR Nashville JUNE 11-15, 2025

Machine learning models typically perform reliably only on inputs drawn from the distribution they were trained on, making Out-of-Distribution (OOD) detection essential for safety-critical applications. While exposing models to example outliers during training is one of the most effective ways to enhance OOD detection, recent studies suggest that synthetically generated outliers can also act as regularizers for deep neural networks. In this paper, we propose an augmentation scheme for synthetic outliers that regularizes a classifier's energy function by adversarially lowering the outliers' energy during training. We demonstrate that our method improves OOD detection performance and adversarial robustness on OOD data on several image classification benchmarks. Additionally, we show that our approach preserves in-distribution generalization.

Background

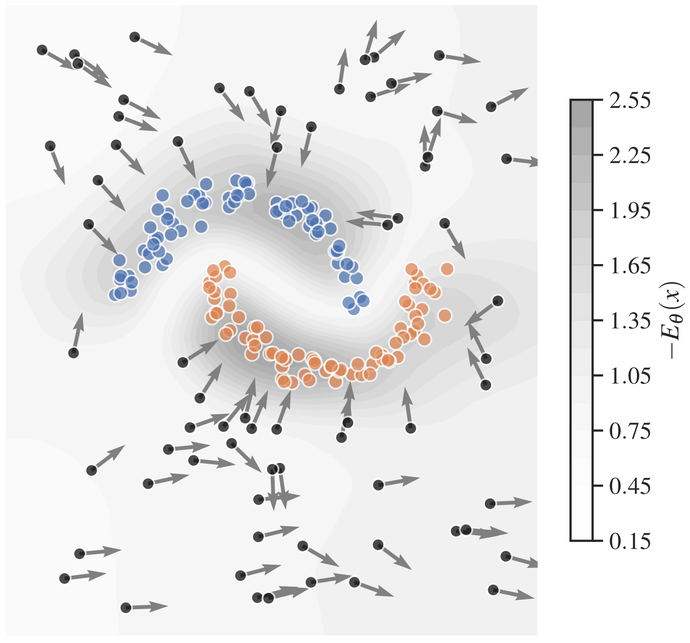
- **Outlier Exposure (OE)**: Train with OOD examples:

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{ID}}}\Big[\mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}),y)\Big] \quad + \lambda\mathbb{E}_{\mathbf{x}'\sim\mathcal{D}_{\text{OOD}}^{\text{train}}}\Big[\mathcal{L}_{\text{OE}}(f_{\boldsymbol{\theta}}(\mathbf{x}'),f_{\boldsymbol{\theta}}(\mathbf{x}),y)\Big]$$

- Outlier Exposure can also be done with outliers sampled from generative models
- **Adversarial Training**: Train with adversarially perturbed in-distribution (ID) examples
- Out-of-Distribution detection: **Energy** E_θ is theoretically aligned with the ID data density:

$$E_\theta(x) = -\log\sum_i \exp(f_\theta(x)_i)$$

- However: Adversarial training has been shown to degrade OOD detection performance and ID classification performance
- Question: Can we use adversarial attacks on (synthetic) outliers during training to tighten the decision boundary?

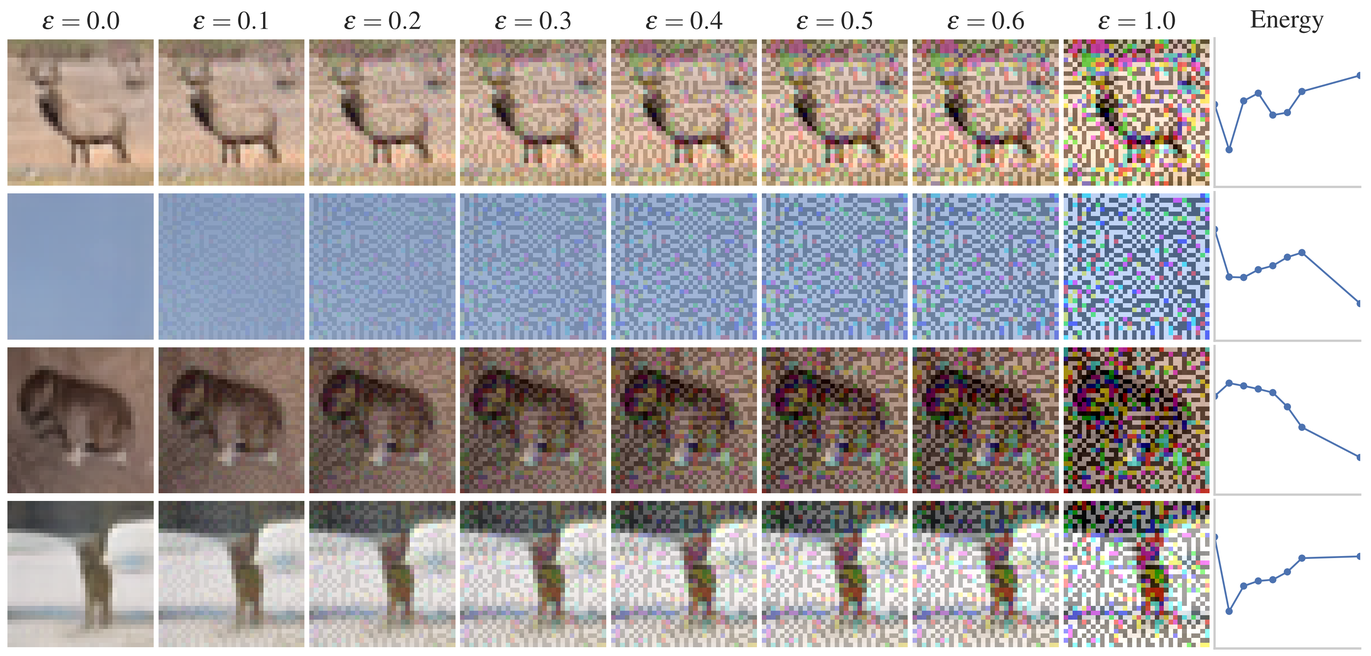


Adversarial Outlier Exposure (AOE)

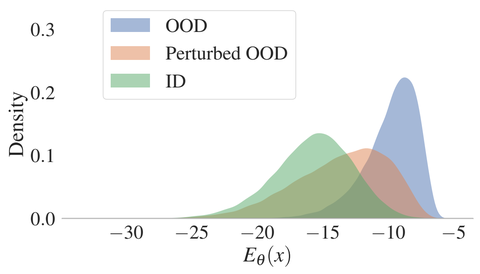
- During training, adversarially augment outliers with perturbations along the model's negative energy gradient
- In particular, we propose FGSM-style augmentations:

$$\hat{\mathbf{x}} \triangleq \mathbf{x} - \epsilon\ \text{sign}(\nabla_{\mathbf{x}} E_{\boldsymbol{\theta}}(\mathbf{x}))$$

- This modifies training outliers such that their energy is more aligned with that of ID data
- Using OE, the model then learns to map these augmented outliers back into high energy regions



Outliers sampled from BigGAN trained on CIFAR-100, with AOE augmentation.



OOD Detection

- AOE improves OOD Detection beyond vanilla Outlier Exposure
- It also outperforms OE augmented with Gaussian noise

Adversarial OOD Detection

- AOE consistently improves robustness to adversarially perturbed OOD data against (FGSM) adversaries targeting the model's energy function

ID Accuracy

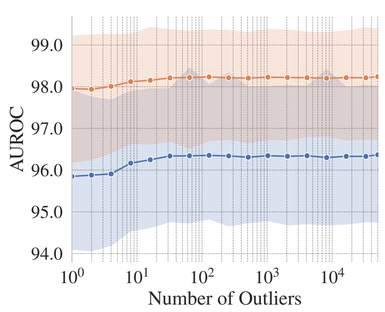
- AOE (mostly) preserves ID classification performance

Convergence

- Stable training, fast convergence

Sample Efficiency

- Additional outliers improve performance, but the effect saturates quickly



Future Work

- Does AOE work with non-synthetic outliers?
- Does AOE scale to larger datasets?
- Why don't stronger adversaries improve results as much?

GitHub