Outlier Synthesis via Hamiltonian Monte Carlo for Out-of-Distribution Detection

Hengzhuang Li, Teng Zhang

School of Computer Science and Technology,

Huazhong University of Science and Technology, Wuhan, China

Contents

- Introduction
- Motivation
- Method
- Results
- Conclusion

Introduction

- Task: fine-tuning based out-of-distribution (OOD) detection without access to auxiliary OOD dataset
- Outlier synthesis: synthesize virtual outliers which serve as surrogated OOD supervision signals
- Results: our proposed framework ***HamOS*** synthesizes high quality outliers and outperforms previous baselines

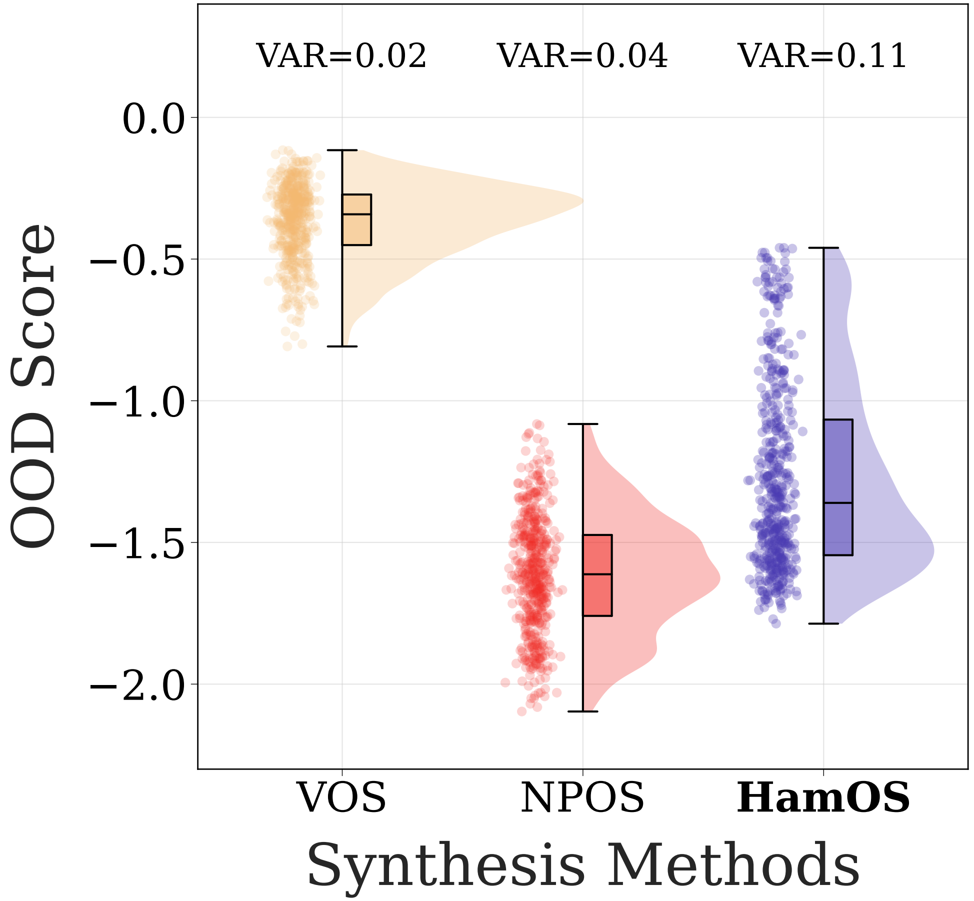
Motivation

- Pixel space outlier synthesis
  - E.g., Dream-OOD [1]
  - To generate pixel space outliers through the generative models, e.g., diffusion model.
- Feature space outlier synthesis
  - E.g., VOS [2], NPOS [3]
  - To generate outliers in the feature space through sampling algorithms, e.g. Gaussian.

[1] Du, et al. Dream the impossible: Outlier imagination with diffusion models. In NIPS, 2023.
[2] Du, et al. Vos: Learning what you don't know by virtual outlier synthesis. In ICLR, 2022.
[3] Tao, et al. Non-parametric outlier synthesis. In ICLR, 2023.

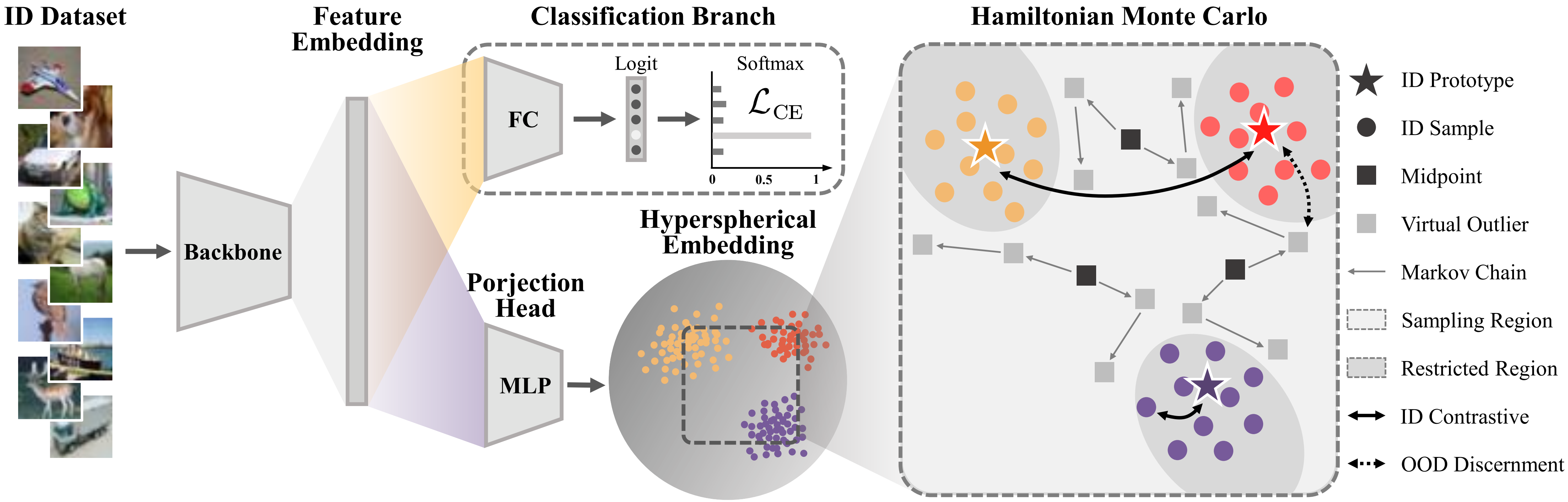
Motivation

- Our goal
  - To efficiently synthesize diverse and representative outliers based solely on the ID data
- Idea
  - Modeling the synthesis process as Markov chain



OOD score distributions

Method



ID Dataset
Feature Embedding
Classification Branch
Hamiltonian Monte Carlo
Backbone
FC
Logit
Softmax
$\mathcal{L}_{CE}$
0
0.5
1
Porjection Head
MLP
Hyperspherical Embedding
ID Prototype
ID Sample
Midpoint
Virtual Outlier
Markov Chain
Sampling Region
Restricted Region
ID Contrastive
OOD Discernment

Method

- Synthesizing Outliers via Hamiltonian Monte Carlo (HMC)
  - Estimating OOD density via the distance to the k-th nearest neighbor: we design a quantitative characterizing of the likelihood that a sample is OOD rather than ID.

$$P^{\mathrm{OOD}}(\boldsymbol{z};\mathcal{Z}_c) = \|\boldsymbol{z}-\boldsymbol{z}_{c(k)}\|_2 \qquad P^{\mathrm{OOD}}(\boldsymbol{z};\mathcal{Z}_u,\mathcal{Z}_v) = \frac{P^{\mathrm{OOD}}(\boldsymbol{z};\mathcal{Z}_u)+P^{\mathrm{OOD}}(\boldsymbol{z};\mathcal{Z}_v)}{2}$$

$$U^{\mathrm{OOD}}(\boldsymbol{z};\mathcal{Z}_u,\mathcal{Z}_v) = -\log P^{\mathrm{OOD}}(\boldsymbol{z};\mathcal{Z}_u,\mathcal{Z}_v) = -\log\sum_{i=u,v} P^{\mathrm{OOD}}(\boldsymbol{z};\mathcal{Z}_i) + \mathrm{constant}$$

  - Synthesizing outliers by OOD density estimation via HMC: we generate virtual outliers along the Markov chains by solving the Hamilton's Equation.

$$H(\boldsymbol{z},\boldsymbol{q}) = U^{\mathrm{OOD}}(\boldsymbol{z}) + \frac{1}{2}\|\boldsymbol{q}\|_2^2$$

  - Rejecting erroneous outliers located within ID clusters: we reject false outliers that conflate with ID embeddings by applying a hard margin according to the ID probability.

$$t_{-} = -\log\max_c P_c^{\mathrm{ID}}(\mathbf{b}_{u,v}) - \delta \quad \mathbf{b}_{u,v} = \frac{\boldsymbol{\mu}_u+\boldsymbol{\mu}_v}{\|\boldsymbol{\mu}_u+\boldsymbol{\mu}_v\|_2}$$

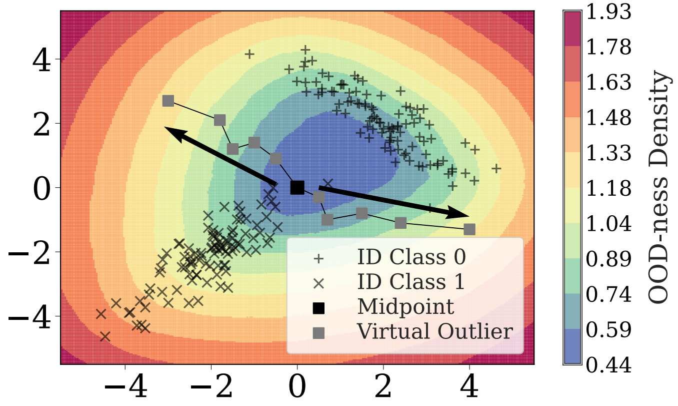
Method

- Training with Synthesized Outliers
  - We fine-tune the model with the OOD discernment loss, the contrastive loss, and the cross-entropy loss to help broaden the gap between ID and OOD data.

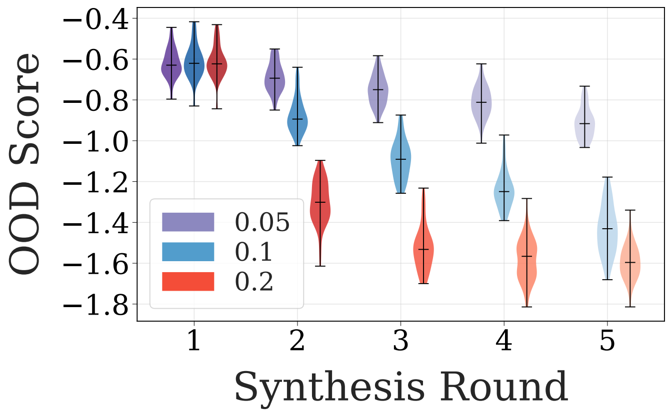
$$\mathcal{L}_{\mathrm{OOD-disc}} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{C}\sum_{j=1}^{C}\log\frac{\exp\left(\boldsymbol{z}_i^{\top}\boldsymbol{\mu}_j/\tau\right)}{\sum_{l=1}^{C}\exp\left(\boldsymbol{z}_i^{\top}\boldsymbol{\mu}_l/\tau\right)}$$

$$\mathcal{L}_{\mathrm{HamOS}} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{ID-con}} + \lambda_d \mathcal{L}_{\mathrm{OOD-disc}}$$

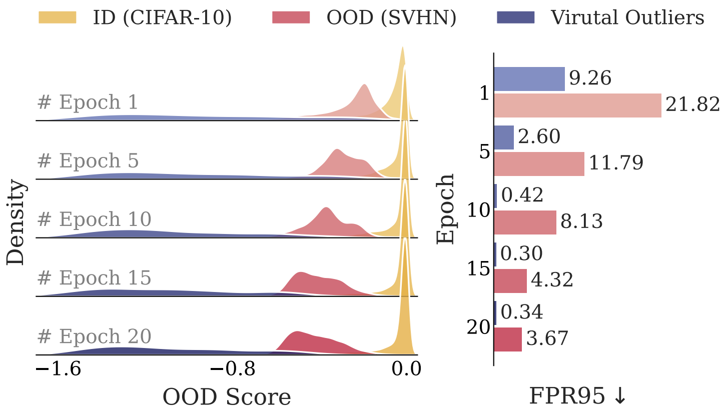
Method



Depiction of the designed OOD-ness density estimation.



Varied OOD scores of the generated outliers at different synthesis rounds.



OOD performance is improved continuously along the training process.

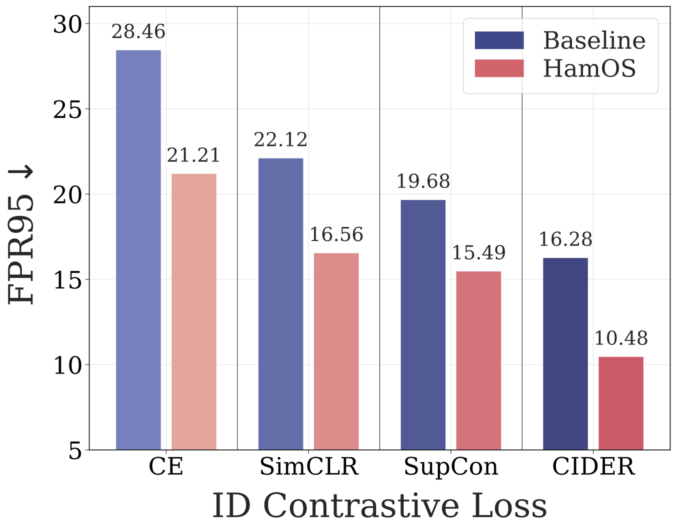
Results

Main results: CIFAR10/100 benchmarks

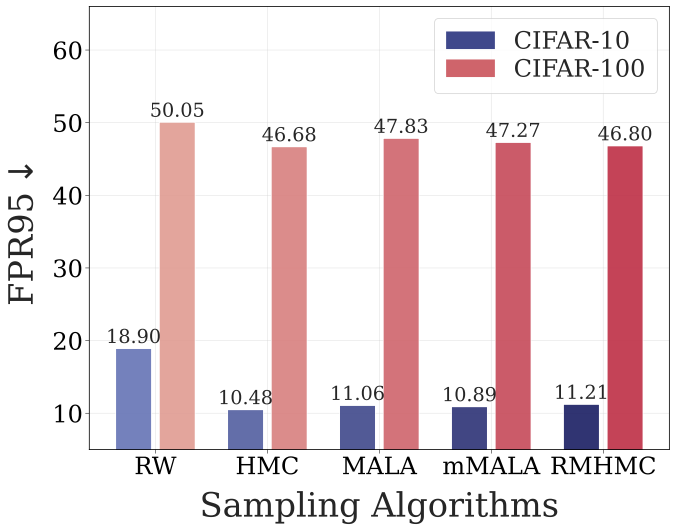
Methods	CIFAR-10				CIFAR-100			
	FPR95↓	AUROC↑	AUPR↑	ID-ACC↑	FPR95↓	AUROC↑	AUPR↑	ID-ACC↑
Post-hoc Methods								
MSP	$32.17_{\pm 6.38}$	$91.10_{\pm 0.71}$	$81.70_{\pm 5.82}$	$\mathbf{95.17_{\pm 0.16}}$	$59.78_{\pm 2.16}$	$77.25_{\pm 1.28}$	$66.86_{\pm 1.58}$	$\mathbf{76.69_{\pm 0.24}}$
ODIN	$58.04_{\pm 18.46}$	$85.70_{\pm 4.17}$	$70.08_{\pm 11.84}$	$\mathbf{95.17_{\pm 0.16}}$	$63.49_{\pm 2.51}$	$78.01_{\pm 1.62}$	$65.20_{\pm 2.19}$	$\mathbf{76.69_{\pm 0.25}}$
EBO	$41.85_{\pm 13.78}$	$91.79_{\pm 1.54}$	$79.70_{\pm 8.10}$	$\mathbf{95.17_{\pm 0.16}}$	$60.86_{\pm 1.87}$	$78.32_{\pm 1.31}$	$66.73_{\pm 1.35}$	$\mathbf{76.69_{\pm 0.24}}$
KNN	$22.86_{\pm 1.12}$	$92.98_{\pm 0.42}$	$88.74_{\pm 0.79}$	$\mathbf{95.17_{\pm 0.16}}$	$56.96_{\pm 2.96}$	$81.01_{\pm 1.19}$	$70.60_{\pm 2.29}$	$\mathbf{76.69_{\pm 0.24}}$
ASH	$54.22_{\pm 26.06}$	$87.37_{\pm 6.60}$	$72.33_{\pm 16.40}$	$95.10_{\pm 0.14}$	$66.84_{\pm 0.87}$	$77.14_{\pm 1.12}$	$62.24_{\pm 0.73}$	$76.20_{\pm 0.23}$
Scale	$63.18_{\pm 23.64}$	$77.74_{\pm 16.24}$	$63.03_{\pm 20.52}$	$95.15_{\pm 0.16}$	$69.27_{\pm 2.31}$	$77.25_{\pm 1.01}$	$61.42_{\pm 1.42}$	$\mathbf{76.69_{\pm 0.24}}$
Relation	$26.28_{\pm 1.63}$	$92.31_{\pm 0.43}$	$86.75_{\pm 0.98}$	$\mathbf{95.17_{\pm 0.16}}$	$59.64_{\pm 2.48}$	$79.69_{\pm 1.08}$	$68.76_{\pm 1.78}$	$\mathbf{76.69_{\pm 0.24}}$
Regularization-based Methods								
CSI	$21.21_{\pm 1.68}$	$93.73_{\pm 0.33}$	$89.74_{\pm 0.68}$	$92.03_{\pm 0.72}$	$69.34_{\pm 0.86}$	$73.46_{\pm 0.37}$	$61.57_{\pm 0.75}$	$61.75_{\pm 0.15}$
SSD+	$18.49_{\pm 1.20}$	$94.85_{\pm 0.57}$	$90.88_{\pm 0.83}$	$93.95_{\pm 0.57}$	$54.03_{\pm 1.92}$	$80.64_{\pm 0.60}$	$69.73_{\pm 1.09}$	$75.63_{\pm 0.39}$
KNN+	$19.68_{\pm 1.86}$	$94.41_{\pm 0.66}$	$90.46_{\pm 0.66}$	$93.79_{\pm 0.63}$	$61.25_{\pm 0.81}$	$78.24_{\pm 0.93}$	$66.64_{\pm 0.88}$	$72.18_{\pm 0.58}$
VOS	$42.37_{\pm 21.13}$	$91.42_{\pm 3.38}$	$79.16_{\pm 11.62}$	$95.05_{\pm 0.05}$	$58.55_{\pm 1.53}$	$81.40_{\pm 0.62}$	$68.33_{\pm 1.61}$	$74.71_{\pm 0.07}$
CIDER	$16.28_{\pm 0.68}$	$95.76_{\pm 0.37}$	$92.36_{\pm 0.06}$	$93.98_{\pm 0.16}$	$49.64_{\pm 1.80}$	$81.77_{\pm 0.95}$	$73.22_{\pm 1.12}$	$75.09_{\pm 0.49}$
NPOS	$14.39_{\pm 0.87}$	$96.61_{\pm 0.26}$	$93.35_{\pm 0.74}$	$93.95_{\pm 0.13}$	$51.41_{\pm 1.88}$	$81.02_{\pm 0.98}$	$72.49_{\pm 1.54}$	$74.53_{\pm 0.62}$
PALM	$32.25_{\pm 4.14}$	$90.54_{\pm 1.46}$	$84.44_{\pm 2.14}$	$93.93_{\pm 0.98}$	$55.13_{\pm 0.97}$	$79.95_{\pm 1.26}$	$70.21_{\pm 1.38}$	$74.67_{\pm 0.36}$
HamOS(ours)	$\mathbf{10.48_{\pm 0.76}}$	$\mathbf{97.11_{\pm 0.26}}$	$\mathbf{94.94_{\pm 0.86}}$	$94.67_{\pm 0.15}$	$\mathbf{46.68_{\pm 1.44}}$	$\mathbf{83.64_{\pm 0.64}}$	$\mathbf{75.52_{\pm 1.30}}$	$76.12_{\pm 0.14}$

Results

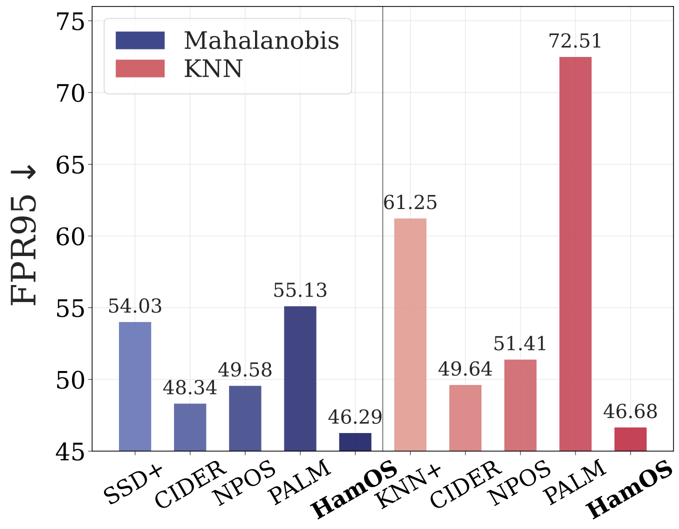
Ablation study



Different contrastive loss



Different variants of HMC



Different scoring functions

Conclusion

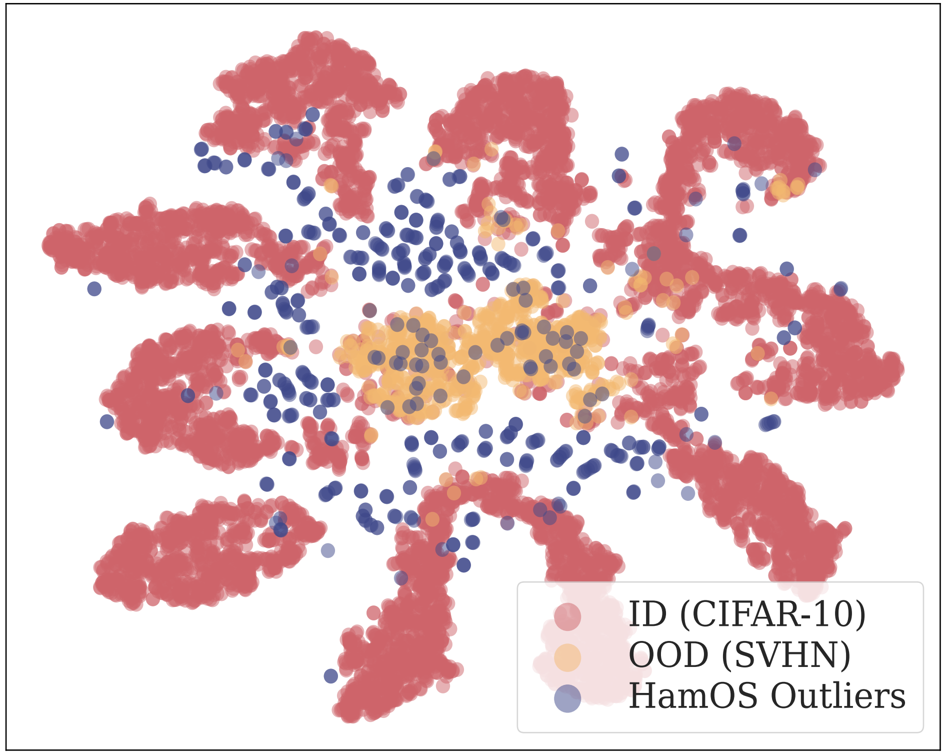
- We propose a novel framework HamOS to synthesize virtual outliers for OOD detection

Please Contact Us !

arXiv

Github

WeChat



Feature visualization via t-SNE