An Information Theoretical View for Out-Of-Distribution Detection

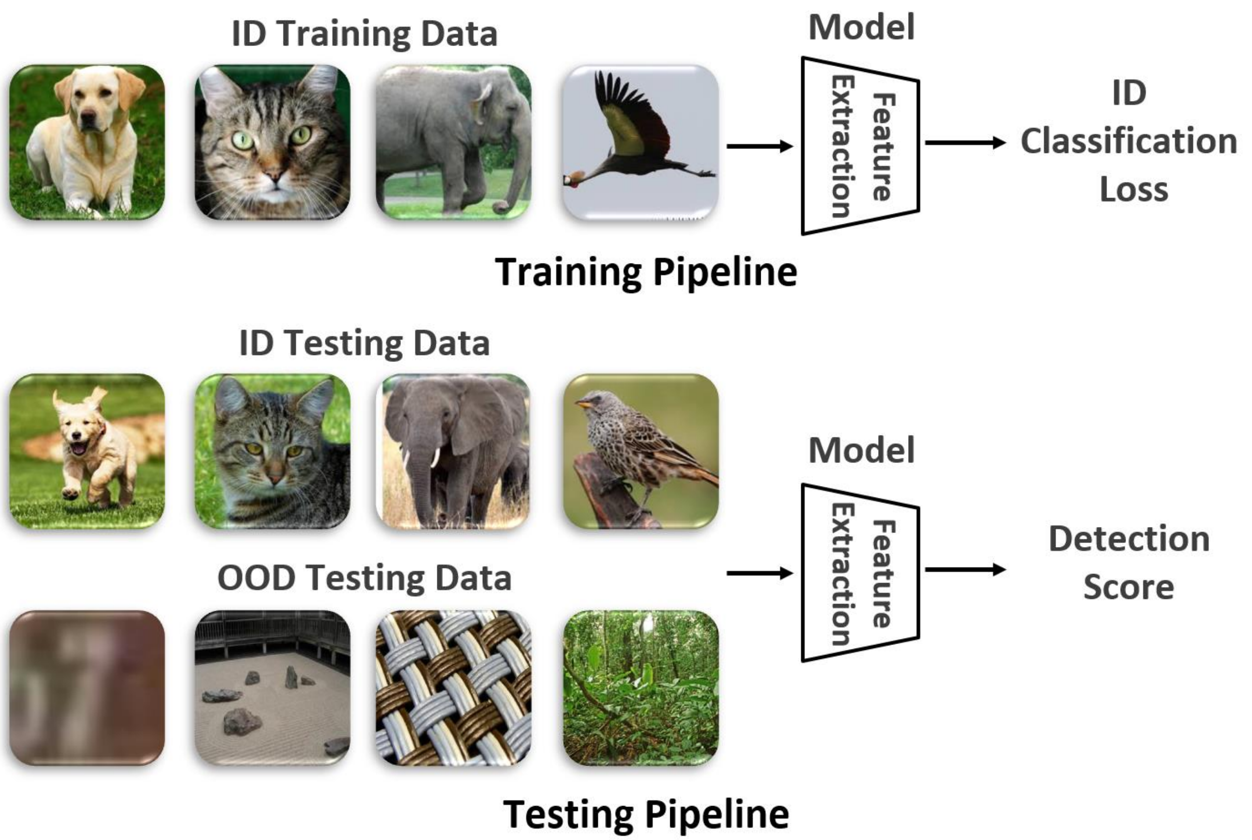
Jinjing Hu[1,2], Wenrui Liu[2,3], Hong Chang[2,3], Bingpeng Ma[3], Shiguang Shan[2,3], Xilin Chen[2,3]

[1]ShanghaiTech University, China
[2]Key Laboratory of AI Safety of CAS, Institute of Computing Technology, Chinese Academy of Sciences (CAS)
[3]University of Chinese Academy of Sciences

Out-Of-Distribution Detection

Out-Of-Distribution(OOD) inputs: samples from an **unknown distribution** that the network has not been exposed to during training phase

Out-Of-Distribution Detection



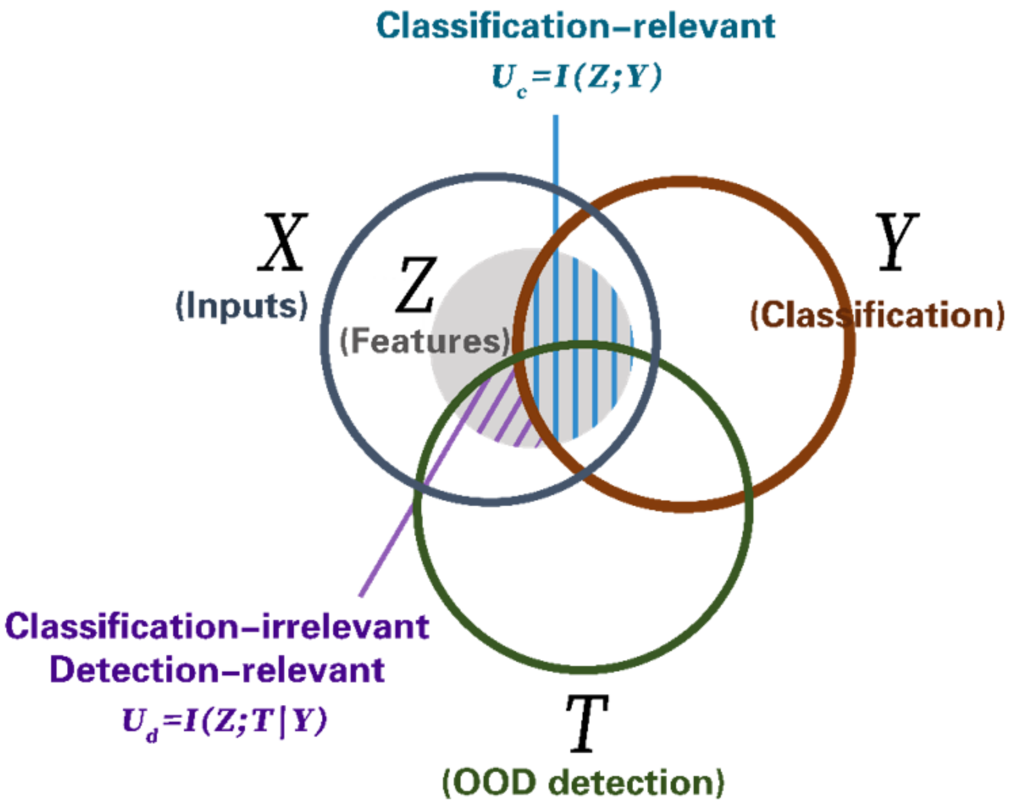
Representation Learning Method: learning **discriminative feature representation** between IDs and OODs.

Information Theory

➢**Information Bottleneck Theory:**

$$\mathcal{L} = I(Z;X) - \beta I(Z;Y)$$

➢ **Illustration:** information relationship between **inputs**, **features**, **classification** and **OOD detection**.



Our Propositions

Proposition 1. *(Over-confidence due to maximizing U_c) Maximizing the mutual information U_c exclusively on ID training data according to Information Bottleneck Theory leads to over-confidence on known classes.*

$$H(Y|Z) \geq H(Y, t = \text{in}|Z) + H(Y, t = \text{out}|Z)$$

Proposition 2. *(Compression of U_d due to optimizing Information Bottleneck theory) Optimizing the classification objective leads to the compression of class-irrelevant detection-relevant information in the representation. Formally, let Z_{min} be the representation variable obtained by optimizing classification objective until convergence. $\forall \epsilon > 0$, we have*

$$I(Z_{min}; T|Y) \leq I(Z_\epsilon; T|Y).$$

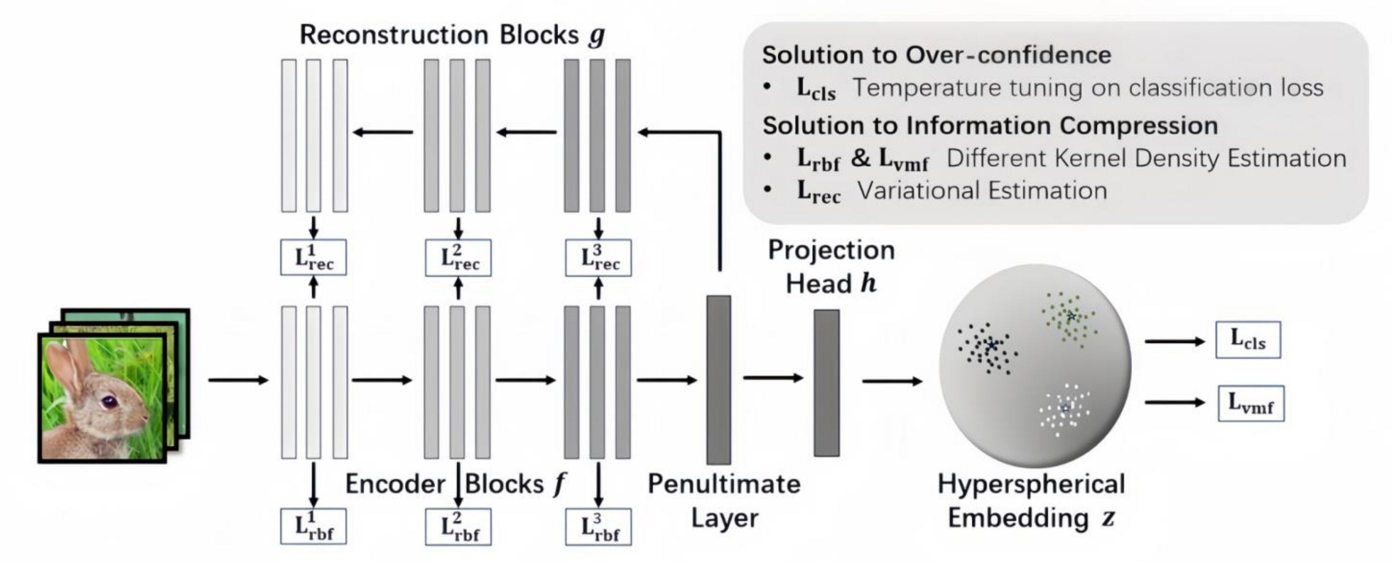
ID classification training formulation can lead to:

Over-confidence on Known Classes (Proposition 1)

Compression of Detection-relevant Information (Proposition 2)

OER Learning Method

➢ Training Procedure:



$$\mathrm{L_{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathbf{z}_i^\top \boldsymbol{\mu}_c/\tau)}{\sum_{j=1}^{C}\exp(\mathbf{z}_i^\top \boldsymbol{\mu}_j/\tau)}$$

$$\mathrm{L_{rbf}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{l}\lambda_l \log\sum_{j\neq i} e^{||f^l(\mathbf{x}_i)-f^l(\mathbf{x}_j)||_2^2}$$

$$\mathrm{L_{vmf}} = \frac{1}{C}\sum_{i=1}^{C}\log\sum_{j\neq i} e^{\boldsymbol{\mu}_i\cdot\boldsymbol{\mu}_j}$$

$$\mathrm{L_{rec}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{l}[||g^l(f(\mathbf{x}_i)) - f^l(\mathbf{x}_i)||_2^2]$$

➢ Inference Procedure:

$$\mathrm{KNN}(\mathbf{z}) = ||\mathbf{z} - \mathbf{z}_{(\mathrm{k})}||_2,$$

Experiments

Main Results

Table 1: OOD detection and ID classification performance on CIFAR-100 (ID) with ResNet-34. ↓ means smaller values are better and ↑ means larger values are better. **Bold** numbers indicate superior results.

Method	SVHN		Places365		LSUN		iSUN		Textures		Average		
	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	ACC↑
MSP	45.2	90.3	84.6	71.8	84.0	74.2	85.7	73.9	81.7	73.2	76.2	76.8	70.3
ODIN	7.8	98.6	79.7	77.3	47.9	92.3	77.3	82.5	70.5	82.5	56.6	86.6	70.3
Maha	87.6	80.7	84.1	73.1	84.3	79.2	84.1	78.7	61.7	84.4	80.3	79.2	70.3
Energy	75.8	77.5	79.1	77.4	41.6	93.1	76.2	82.7	68.3	82.9	68.2	82.7	70.3
DICE	43.7	97.2	85.0	75.9	43.7	95.7	75.2	80.9	75.0	89.8	64.5	87.9	70.3
VOS	77.4	74.1	80.8	74.5	75.6	82.6	68.3	85.4	61.5	85.3	72.8	80.3	74.3
SSD+	40.4	94.1	79.8	78.9	50.9	91.7	81.1	83.3	54.6	89.6	61.4	87.3	**75.9**
KNN+	45.7	91.1	79.5	**79.3**	48.5	91.0	77.4	82.4	53.5	88.8	60.9	86.1	75.9
NPOS	15.4	96.8	79.3	71.3	43.2	87.4	47.7	86.4	45.2	89.4	46.1	86.2	75.5
CIDER	16.1	97.6	**78.3**	75.1	17.1	96.2	49.5	89.2	36.4	92.0	39.4	90.0	75.1
OER	**6.1**	**98.3**	80.3	70.9	**14.9**	**96.1**	**23.2**	**95.9**	**17.8**	**95.1**	**28.4**	**91.2**	74.6

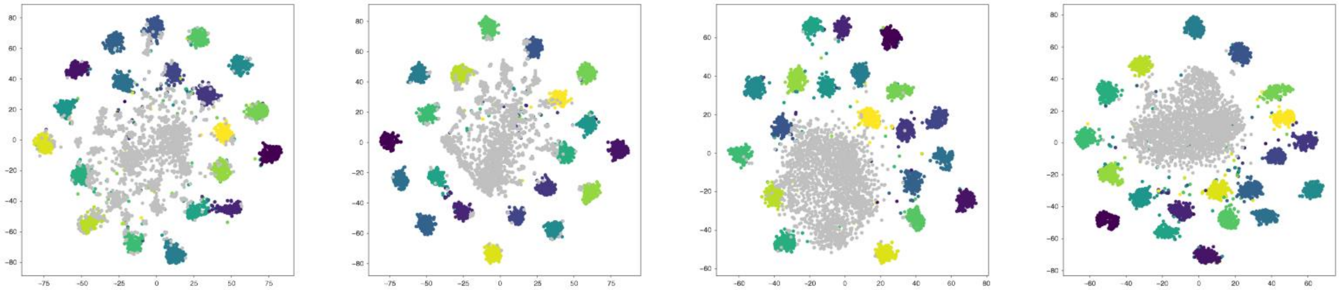
Ablation of Regularization Losses

Table 2: Ablation of proposed loss functions on different ID datasets.

L_{cls}	L_{vmf}	L_{rec}	L_{rbf}	CIFAR-100		ImageNet-100	
				FPR95↓	AUROC↑	FPR95↓	AUROC↑
✓				60.8	85.3	54.8	88.0
✓	✓			39.4	90.0	51.3	88.4
✓	✓	✓		34.7	90.5	51.1	88.6
✓	✓		✓	38.4	90.2	46.2	90.4
✓	✓	✓	✓	**28.4**	**91.2**	**43.7**	**90.9**

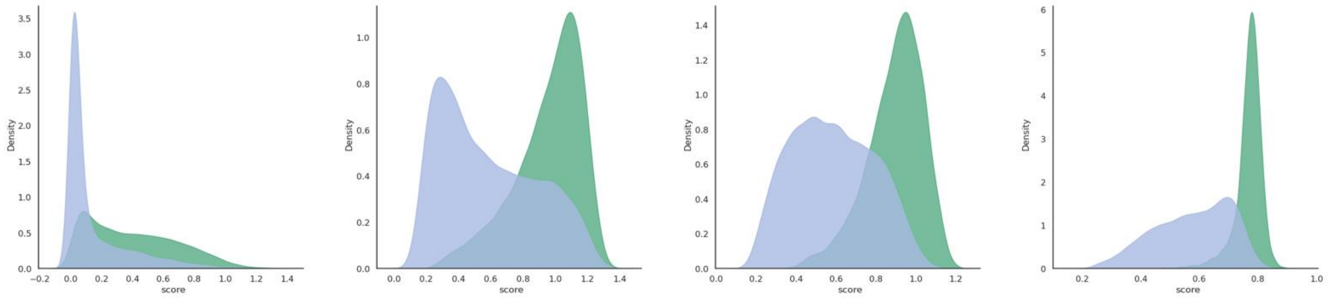
Visualizations

◻ T-SNE Visualization of Feature Distribution



(a) CE (b) SSD+ (c) CIDER (d) OER(ours)

◻ Visualization of OOD Score Distribution



(e) CE (f) SSD+ (g) CIDER (h) OER(ours)

OER Enhances the **Separability** between IDs and OODs.

Conclusion

- ID classification formulation can lead to **over-confidence** and **undesired compression of OOD detection-relevant information**.

- OER could decrease model's confidence based on **temperature coefficient tuning**, and **increase the mutual information** between feature representation and potential OODs.

- OER could effectively **enhance OOD detection** without **compromising ID classification accuracy**.