STEP: Out-of-distribution Detection in the Presence of Limited In-distribution Labeled Data

Zhi Zhou[1*], Lan-Zhe Guo[1*], Zhanzhan Cheng[2], Yu-Feng Li[1†], Shiliang Pu[2]

1 Nanjing University, Nanjing, China
2 Hikvision Research Institute, Hangzhou, China

Learning And Mining from DatA
http://www.lamda.nju.edu.cn

† Corresponding author
* Contribute to this work equally

What is this work about

Learning And Mining from DatA
http://www.lamda.nju.edu.cn

OOD detection protects the **safety** of neural networks **in real-world applications**.

However, OOD detection in a semi-supervised fashion is underexplored, which challenges in the following two aspects:

- **Labeled data is insufficient**
- **Unlabeled data is mixed with both in- and out-of-distribution samples**

These two points meet the situation in real-world semi-supervised problems.

✓ In our work, we summarize a novel and practical **semi-supervised out-of-distribution detection setting** and propose a STEP approach for this setting.

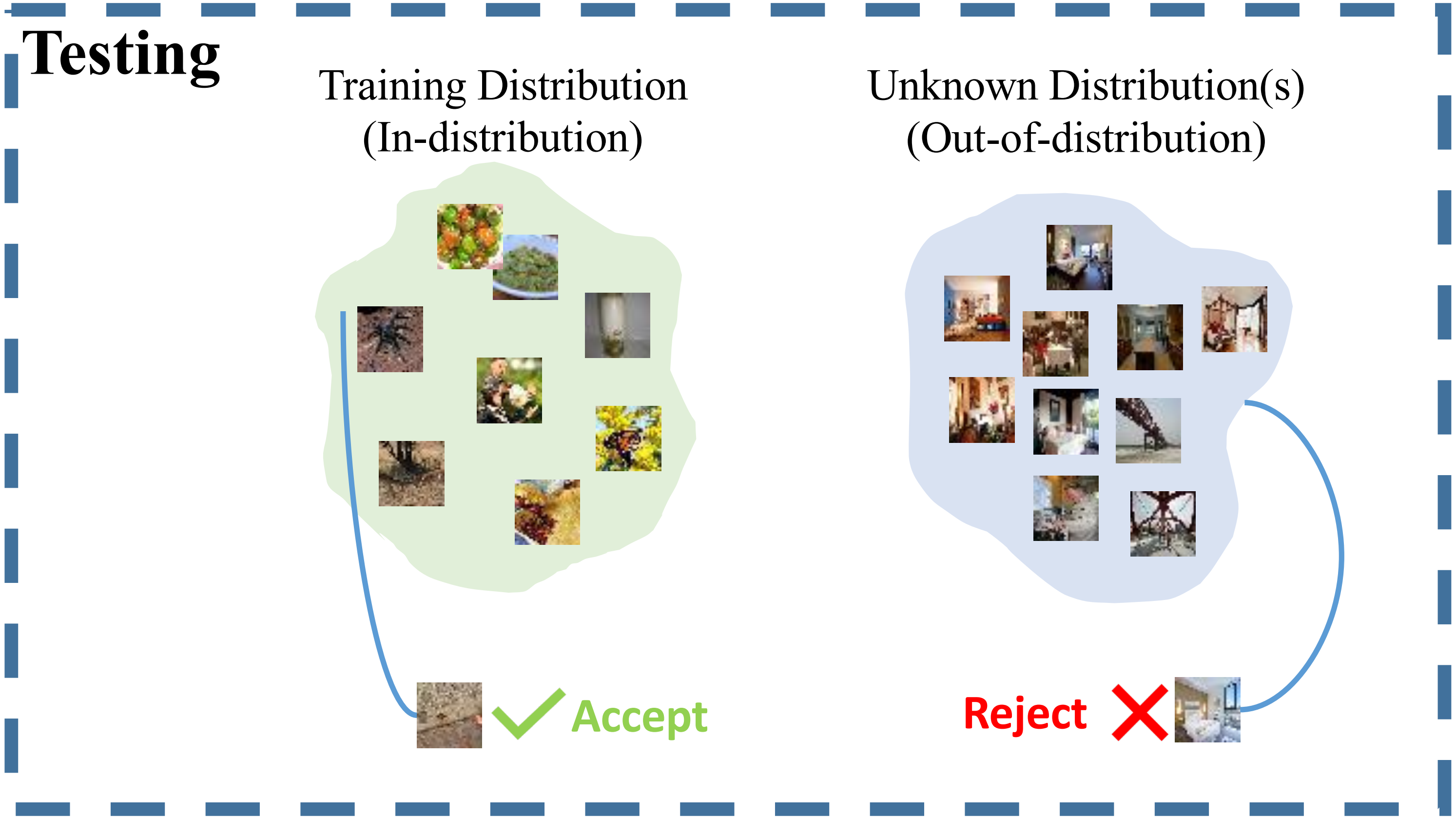
✓ Our proposal is **clearly better than** **two baselines** and **the SOTA out-of-distribution detection method** evaluated by 4 **metrics** on **8 benchmark data sets**.

Outline

- **Motivation**
- STEP Approach
- Experiments
- Conclusions

OOD Detection

OOD Detection: Decide whether a test sample is drawn from training data distribution or not.

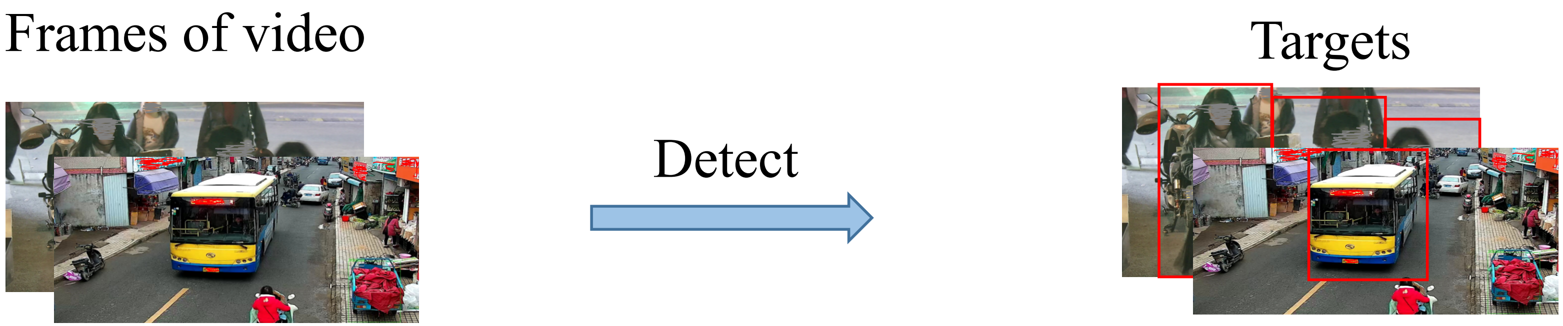


Real Situations

In many real-world applications:

- ✓ Labeled data is limited while the others remain unlabeled.
- ✓ Unlabeled data is mixed with both in- and out-of-distribution samples.

For example: **Video Security System in Hikvision**

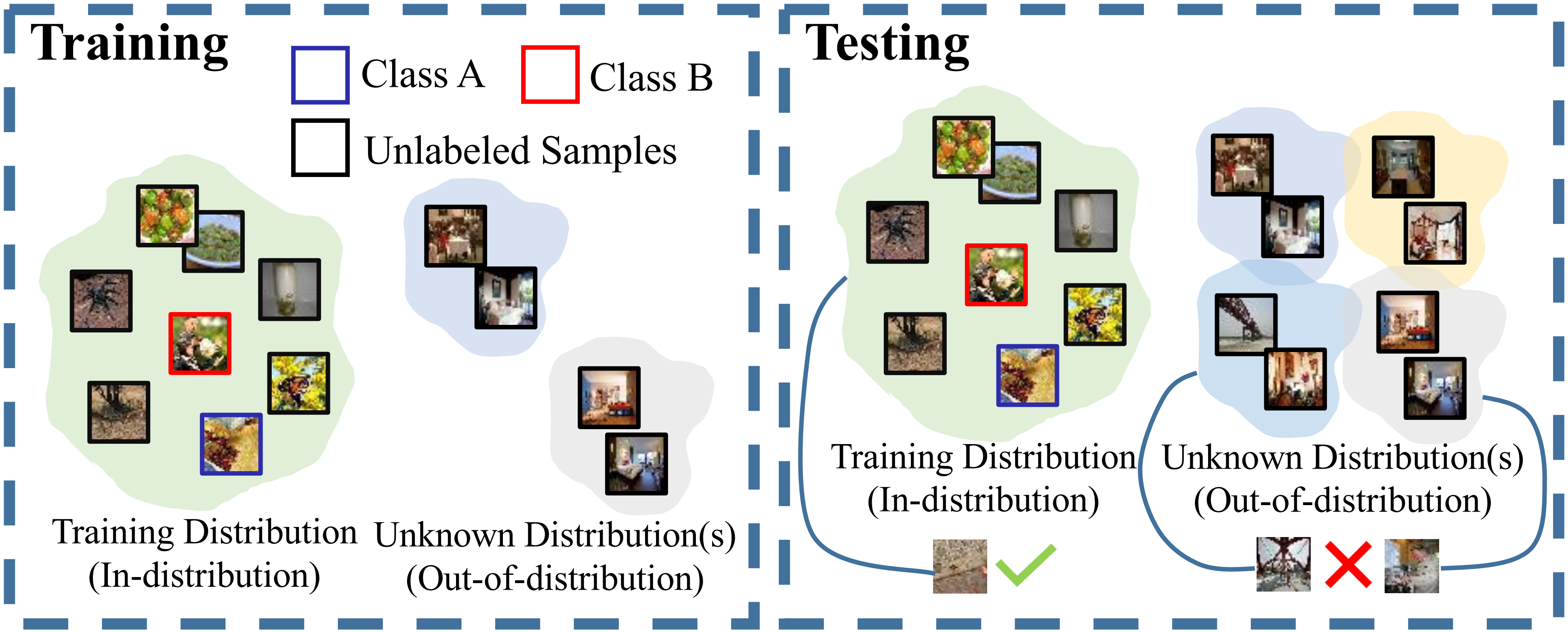


- ➢ Label is difficult to obtain as it requires manual verification. (Labeled data is limited)
- ➢ Millions of videos are generated every. Not all are guaranteed to be relevant. (Unlabeled data contains OOD samples)
- ➢ In different environments(e.g., foggy, sandy), the accuracy of the system is severely affected. (Detecting OOD samples is necessary)

Semi-supervised OOD Detection

Setting:

- Limited in-distribution labeled data
- Large amounts of unlabeled data drawn from both in- and out-of-distribution
- Detecting OOD samples from both known unlabeled data and unknown testing data



Outline

- Motivation
- **STEP Approach**
- Experiments
- Conclusions

STEP Approach

A classical OOD detection method: Mahalanobis Distance

- Mahalanobis Distance between two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is defined as:

$$\mathcal{MD}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\top} \widehat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)}$$

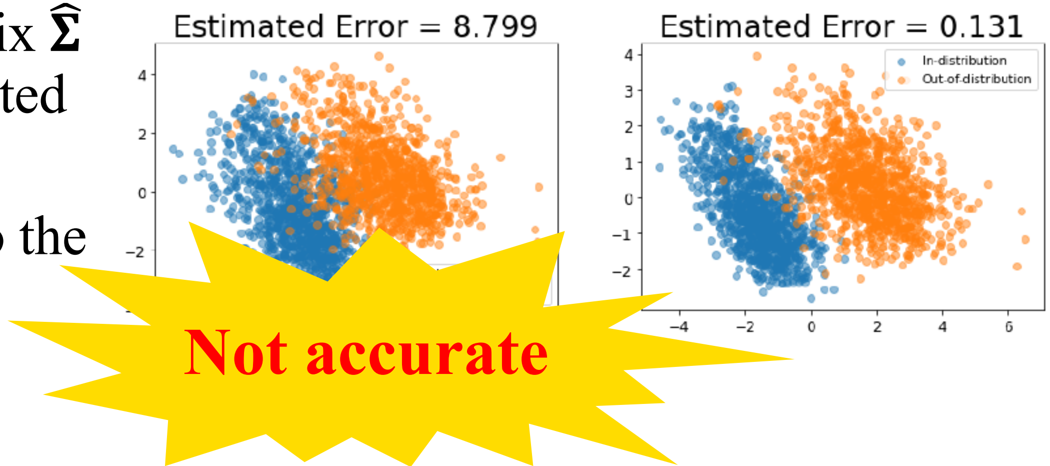
where $\widehat{\boldsymbol{\Sigma}}$ is the covariance matrix estimated on all in-distribution samples.

- The OOD detection confidence score of a testing sample $\boldsymbol{x}$ is defined as:

$$Score_{\mathcal{MD}}(\boldsymbol{x}) = \min_{c \in \{c_1, c_2, \ldots, c_k\}} \mathcal{MD}(\boldsymbol{x}, \boldsymbol{\mu}_c)$$

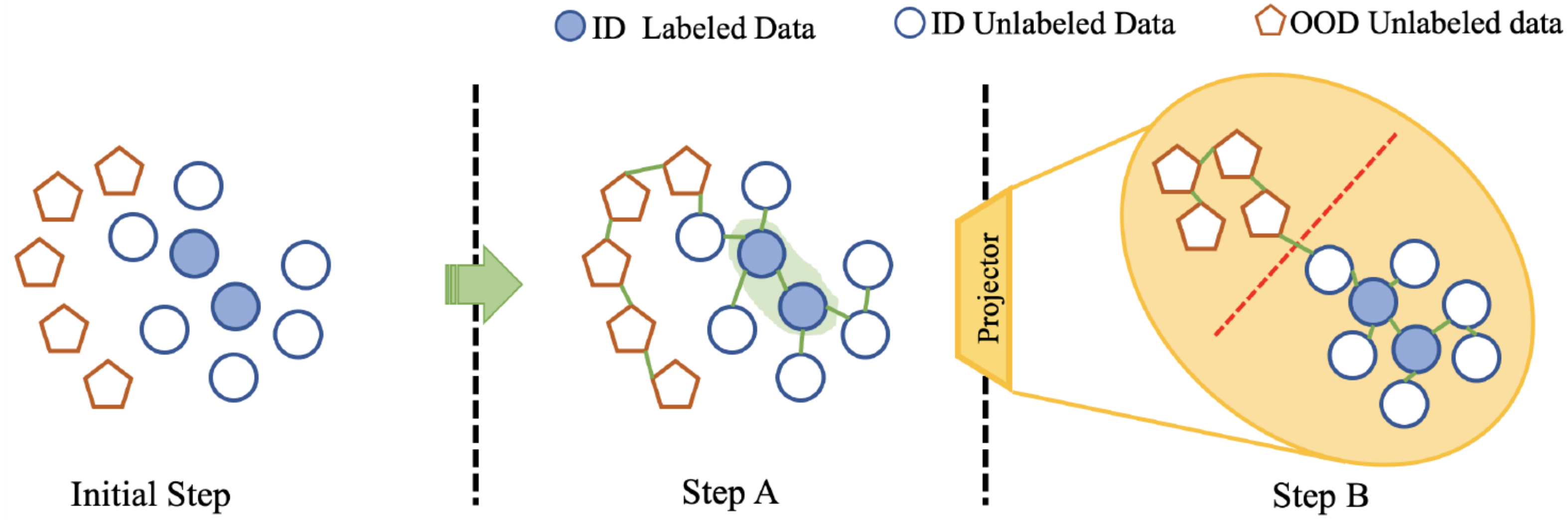
where $\boldsymbol{\mu}_c$ denotes the center of samples that belong to class c.

- However, the covariance matrix $\widehat{\boldsymbol{\Sigma}}$ is hard to be accurately estimated in our setting.
- Large estimation error leads to the degradation of OOD detection performance.



STEP Approach

Learning to project samples into space where a large margin separates ID samples and OOD samples.



➢ Inspired by the topological technology and cluster assumption, we want to project the sample into a space that satisfies the following constraints:

$$\begin{aligned} \max_{\mathbf{P}} \quad & \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_l \cup \mathcal{D}_u} \|\mathbf{P}\mathbf{x}_i - \mathbf{P}\mathbf{x}_j\|_2 \\ \text{s.t.} \quad & \|\mathbf{P}\mathbf{x}_i - \mathbf{P}\mathbf{x}_n\|_2 = \mathcal{MD}(\mathbf{x}_i, \mathbf{x}_n), \\ \text{if} \quad & \mathbf{x}_n \in \mathcal{B}_k(\mathbf{x}_i) \end{aligned}$$

where $\mathcal{B}_k(\boldsymbol{x}_i)$ is the set of k nearest neighbours of $\boldsymbol{x}_i$.

STEP Approach

- We define L_{Keep} and L_{Unzip} that can be directly optimized to approximately achieve our objective:

$$\begin{cases} L_{Keep} & = \max(0, \|\mathbf{P}\mathbf{x}_i - \mathbf{P}\mathbf{x}_n\|_2 - \mathcal{MD}(\mathbf{x}_i, \mathbf{x}_n)), \\ L_{Unzip} & = -\|\mathbf{P}\mathbf{x}_i - \mathbf{P}\mathbf{x}_j\|_2. \end{cases}$$

- Minimum L2 distance can be directly used as the confidence score:

$$\mathcal{N}(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{P}\mathbf{x}_i - \mathbf{P}\mathbf{x}_j\|_2$$

$$Score(\mathbf{x}) = \min_{c \in \{c_1, c_2, \ldots, c_K\}} \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_c)$$

where $\boldsymbol{\mu}_c$ denotes the center of samples which belong to class c.

Outline

Learning And Mining from DatA
http://www.lamda.nju.edu.cn

- Motivation
- STEP Approach
- **Experiments**
- Conclusions

Setup

8 benchmark data sets

- 2 In-distribution data sets:
  - **CIFAR-10**
  - **CIFAR-100**
- 4 Out-of-distribution data sets:
  - **TINc, TINr**
  - **LSUNc, LSUNc**

5 metrics

- ✓ AUROC
- ✓ FPR at 95% TPR
- ✓ Detection Error
- ✓ AUPR-In
- ✓ AUPR-Out

Compared Methods

- ☐ ODIN [Liang et al., ICLR 2018]
- ☐ Mahalanobis [Lee et al., NeurIPS 2018]
- ☐ Unsupervised OOD Detection [Yu et al., ICCV, 2019]

Results on Benchmarks

Experiment results evaluated by AUROC and FPR.

Metrics	ID Dataset	OOD Dataset	ODIN	MAH †	UOOD	UOOD †	STEP
AUROC ↑	Cifar10	TINc	81.00 ± 6.30	87.67 ± 2.47	90.46 ± 9.74	99.07 ± 0.48	**99.99 ± 0.00**
		TINr	59.10 ± 2.08	86.88 ± 0.87	84.67 ± 9.41	92.63 ± 3.42	**95.61 ± 0.36**
		LSUNc	76.17 ± 5.37	97.68 ± 0.09	96.92 ± 2.04	98.79 ± 0.67	**99.99 ± 0.00**
		LSUNr	69.05 ± 3.49	90.41 ± 1.00	80.87 ± 24.45	97.81 ± 0.94	**99.07 ± 0.20**
	Cifar100	TINc	61.65 ± 6.71	71.15 ± 2.20	98.34 ± 1.57	98.84 ± 0.83	**99.99 ± 0.01**
		TINr	54.46 ± 0.74	73.94 ± 1.79	84.80 ± 8.87	**95.31 ± 0.93**	93.51 ± 1.17
		LSUNc	46.99 ± 4.99	93.91 ± 3.41	97.49 ± 1.48	99.31 ± 0.62	**99.99 ± 0.00**
		LSUNr	52.06 ± 2.24	78.45 ± 1.11	97.61 ± 0.55	**98.96 ± 0.40**	98.20 ± 0.56
FPR at 95%TPR ↓	Cifar10	TINc	53.37 ± 10.55	44.17 ± 6.43	29.35 ± 30.05	2.75 ± 1.65	**0.00 ± 0.00**
		TINr	89.76 ± 1.45	58.57 ± 3.09	31.72 ± 11.50	19.61 ± 9.50	**17.63 ± 1.10**
		LSUNc	64.06 ± 9.12	7.73 ± 0.46	6.59 ± 3.22	3.56 ± 1.93	**0.00 ± 0.00**
		LSUNr	76.89 ± 5.04	45.41 ± 3.87	32.69 ± 31.93	6.49 ± 2.89	**4.48 ± 1.02**
	Cifar100	TINc	84.24 ± 8.02	90.15 ± 1.99	5.22 ± 5.59	3.16 ± 2.25	**0.00 ± 0.01**
		TINr	90.10 ± 0.46	80.55 ± 1.89	29.09 ± 15.68	**11.10 ± 4.21**	23.21 ± 4.14
		LSUNc	93.49 ± 2.42	24.93 ± 21.75	6.24 ± 3.80	1.93 ± 2.43	**0.00 ± 0.00**
		LSUNr	89.79 ± 0.79	69.69 ± 2.42	4.92 ± 1.33	**2.39 ± 0.74**	8.25 ± 3.14

STEP gives **the best result** on benchmark datasets, and still give **competitive results** even if the result is not the best.

Results on Benchmarks

http://www.lamda.nju.edu.cn

Experiment results evaluated by Detection Error, AUPR-In and AURP-Out.

Detection Error ↓	Cifar10	TINc	25.53 ± 4.67	19.93 ± 2.63	11.59 ± 11.35	2.54 ± 1.27	**0.12 ± 0.01**
		TINr	43.04 ± 1.48	20.14 ± 0.82	18.07 ± 5.55	11.71 ± 4.56	**10.77 ± 0.52**
		LSUNc	29.57 ± 3.82	6.28 ± 0.25	4.20 ± 2.12	2.58 ± 1.32	**0.11 ± 0.01**
		LSUNr	35.52 ± 2.46	16.23 ± 0.95	18.40 ± 15.68	4.99 ± 1.91	**4.66 ± 0.57**
	Cifar100	TINc	40.95 ± 5.07	32.58 ± 1.64	3.67 ± 3.62	2.76 ± 1.00	**0.32 ± 0.06**
		TINr	46.36 ± 0.56	31.09 ± 1.44	16.53 ± 7.87	**6.88 ± 2.33**	13.26 ± 1.61
		LSUNc	48.47 ± 1.61	11.20 ± 3.73	4.24 ± 2.34	2.06 ± 1.54	**0.23 ± 0.04**
		LSUNr	46.73 ± 0.66	27.33 ± 1.03	3.11 ± 0.78	**1.90 ± 0.51**	6.40 ± 1.32
AUPR-In ↑	Cifar10	TINc	76.80 ± 8.20	85.35 ± 2.86	89.31 ± 10.05	98.59 ± 0.67	**99.99 ± 0.00**
		TINr	57.10 ± 2.11	86.79 ± 1.17	79.02 ± 12.17	88.72 ± 4.93	**94.71 ± 0.51**
		LSUNc	72.16 ± 6.60	96.70 ± 0.21	94.78 ± 4.07	98.31 ± 0.92	**100.00 ± 0.00**
		LSUNr	65.37 ± 3.39	89.93 ± 1.23	79.41 ± 19.89	96.86 ± 1.27	**99.02 ± 0.20**
	Cifar100	TINc	58.29 ± 5.01	71.18 ± 2.69	97.55 ± 2.04	98.24 ± 1.50	**99.99 ± 0.01**
		TINr	52.96 ± 0.59	70.95 ± 2.20	77.32 ± 9.81	91.67 ± 1.29	**91.91 ± 1.34**
		LSUNc	47.41 ± 2.86	92.26 ± 2.17	95.45 ± 2.32	99.09 ± 0.88	**99.99 ± 0.00**
		LSUNr	50.47 ± 1.75	74.22 ± 1.14	95.53 ± 0.95	**98.11 ± 0.78**	98.07 ± 0.52
AUPR-Out ↑	Cifar10	TINc	83.63 ± 5.11	88.67 ± 2.28	91.34 ± 8.69	99.32 ± 0.35	**99.99 ± 0.00**
		TINr	58.83 ± 1.77	84.26 ± 0.95	89.21 ± 6.22	94.60 ± 2.70	**96.31 ± 0.28**
		LSUNc	78.43 ± 5.12	98.16 ± 0.12	98.01 ± 1.18	99.14 ± 0.48	**99.99 ± 0.00**
		LSUNr	70.51 ± 3.97	88.84 ± 1.20	84.45 ± 21.48	98.41 ± 0.70	**99.14 ± 0.19**
	Cifar100	TINc	62.88 ± 7.90	65.14 ± 2.21	98.77 ± 1.23	99.08 ± 0.51	**99.99 ± 0.01**
		TINr	55.94 ± 0.71	71.57 ± 1.71	89.44 ± 6.96	**96.84 ± 0.82**	94.66 ± 1.07
		LSUNc	49.91 ± 4.42	93.77 ± 5.30	98.33 ± 0.99	99.39 ± 0.48	**99.99 ± 0.00**
		LSUNr	55.18 ± 1.56	78.19 ± 1.33	98.49 ± 0.37	**99.32 ± 0.24**	98.35 ± 0.56

Other Results

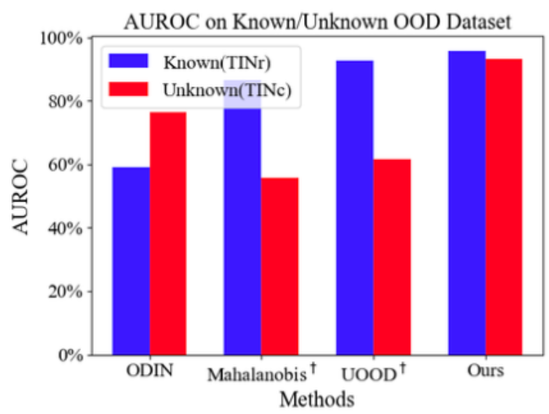
Ablation Study

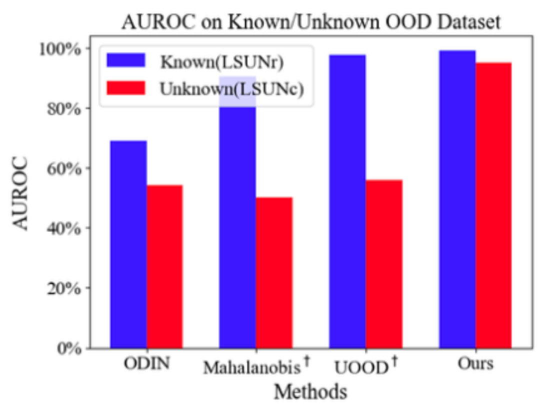
Different parts of STEP				Data set pair	
MAH	KNN	Unzipping	Sturture-Keep	Cifar10-TINr	Cifar10-LSUNr
✓				90.96 ± 0.28	93.46 ± 0.51
✓	✓			91.26 ± 1.74	97.35 ± 0.45
✓	✓	✓		79.58 ± 0.69	80.38 ± 0.95
✓	✓	✓	✓	95.62 ± 0.39	99.07 ± 0.20

The four components proposed in this STEP can only get the best results if they are integrated.

Generalization of OOD Detection

The STEP approach gives a very high and relatively close performance on both known and unknown OOD data sets, which shows strong generalization.





Other experiments can be found in our paper and supplementary materials.

Outline

Learning And Mining from DatA
http://www.lamda.nju.edu.cn

- Motivation
- STEP Approach
- Experiments
- **Conclusions**

Conclusions

Learning And Mining from DatA
http://www.lamda.nju.edu.cn

In this paper, we consider a **novel** and **realistic** setting:

Semi-supervised Out-of-distribution Detection

- ✓ A novel OOD detection setting with realistic applications
- ✓ A simple yet effective STEP approach
- ✓ Extensive experiments demonstrate the effectiveness of STEP

Future work

- ➢ Imbalances problems may emerge in real applications

Code:

https://github.com/WNJXYK/Step

Thank you!

If you are interested in, feel free to contact me:

Zhi Zhou (zhouz@lamda.nju.edu.cn)