西安電子科技大學
XIDIAN UNIVERSITY

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Discriminating Known from Unknown Objects via Structure-Enhanced Recurrent Variational AutoEncoder

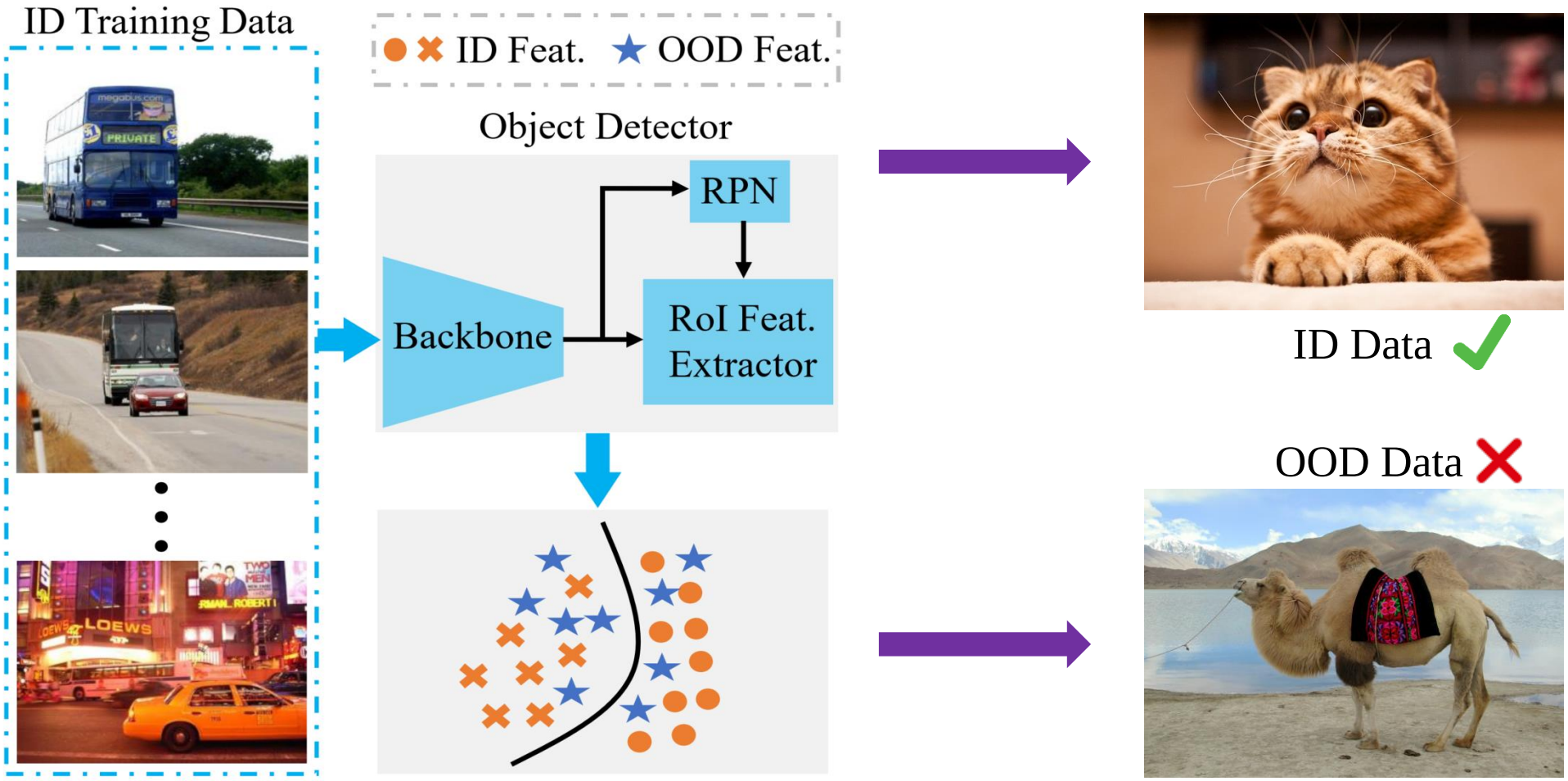
Aming Wu, Cheng Deng

School of Electronic Engineering, Xidian University, Xi'an, China

amwu@xidian.edu.cn, chdeng@mail.xidian.edu.cn

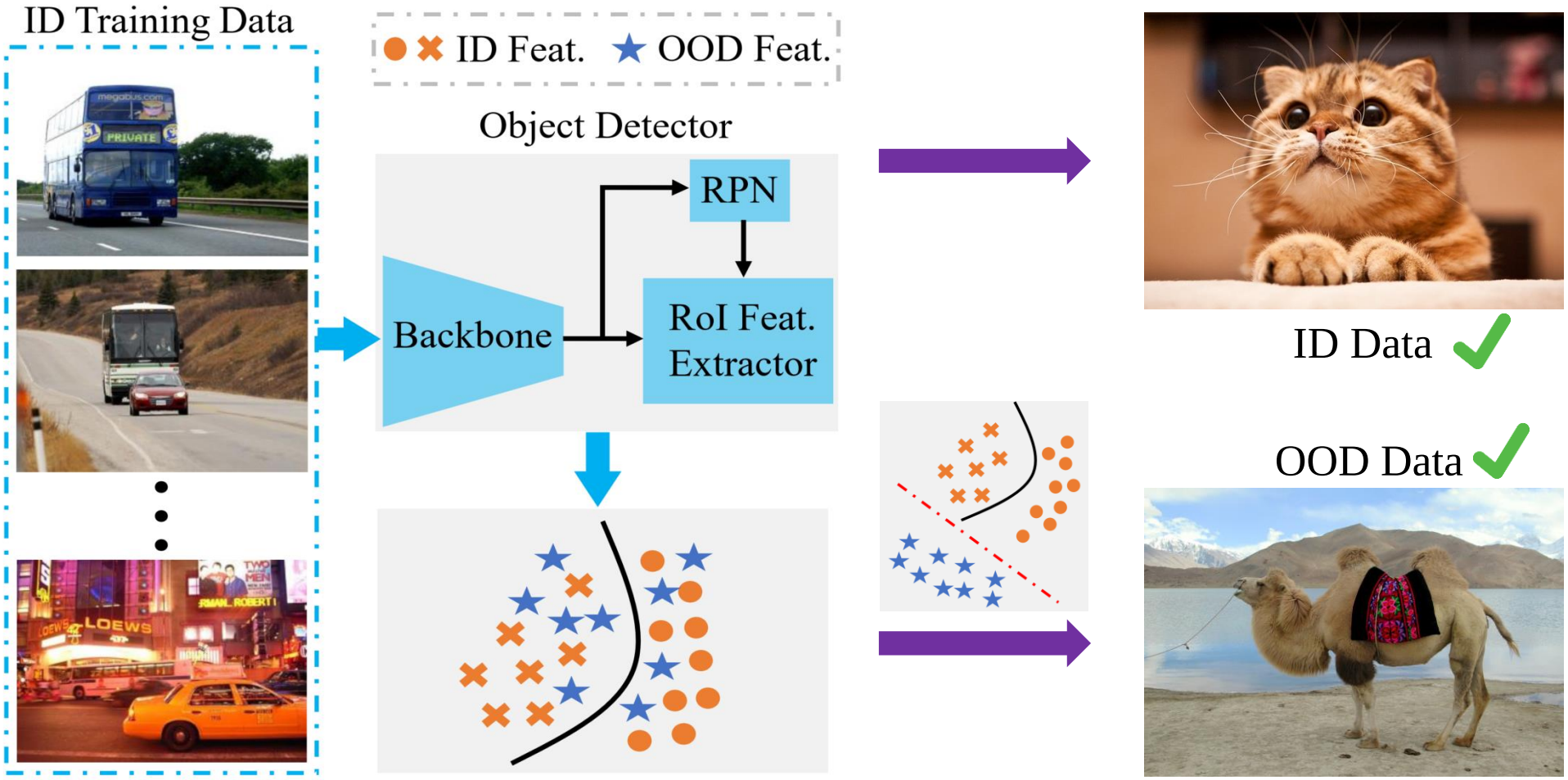
Unsupervised Out-of-Distribution Object Detection (OOD-OD)

- Unsupervised Out-of-Distribution Object Detection (OOD-OD) aims to detect the objects never-seen-before during training without accessing any auxiliary data



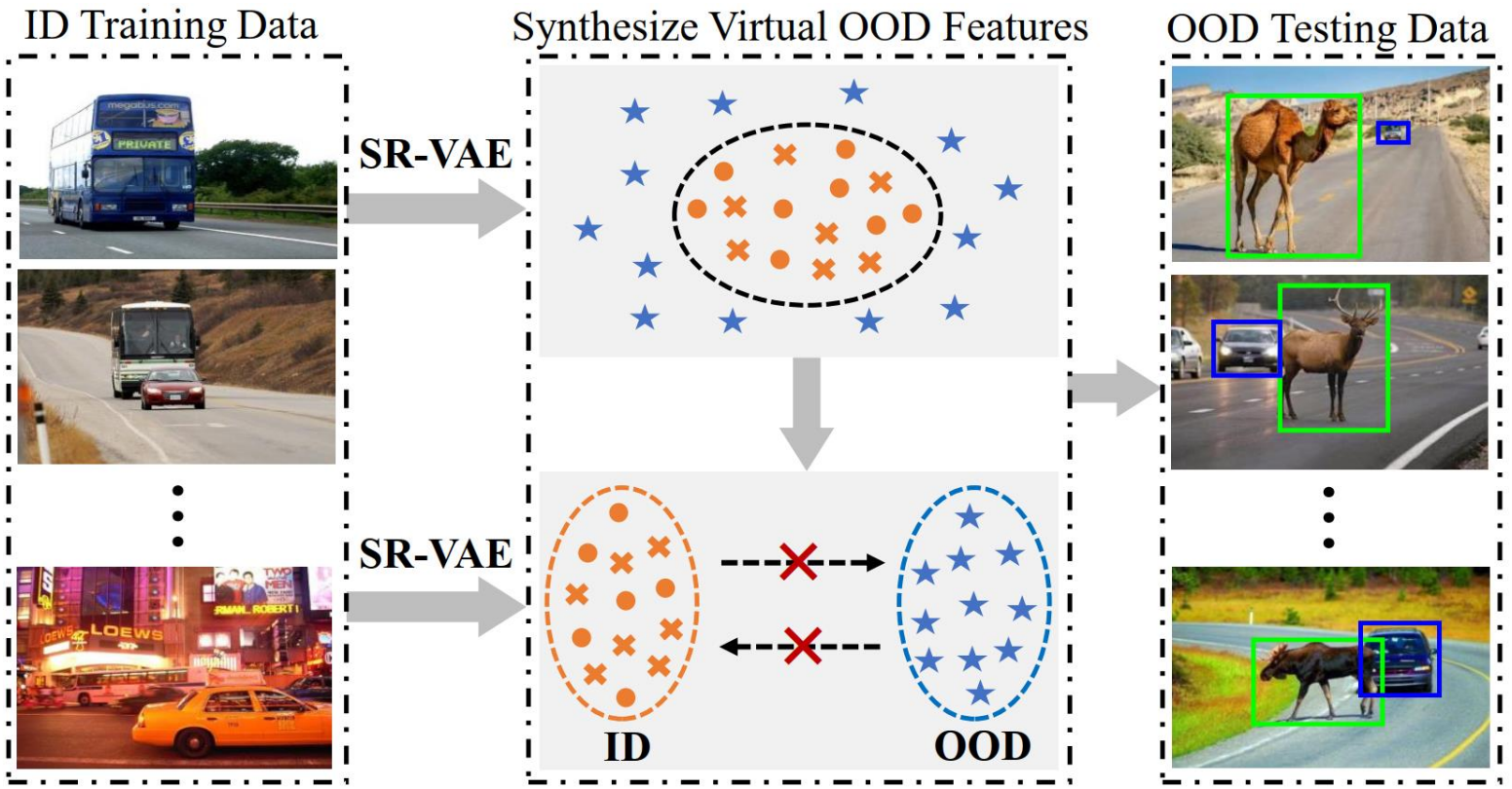
Unsupervised Out-of-Distribution Object Detection (OOD-OD)

- For unsupervised OOD-OD, since there is no auxiliary data available for supervision, leveraging the known in-distribution (ID) data to enhance the detector's discrimination ability becomes the critical challenge



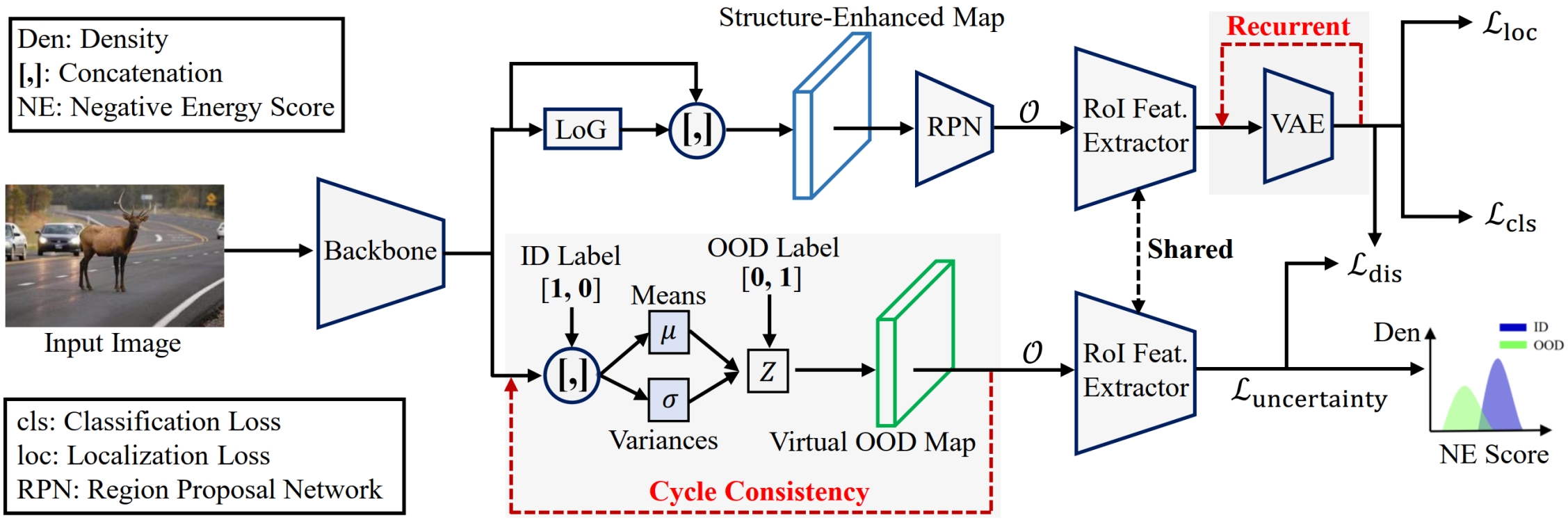
Structure-Enhanced Recurrent Variational AutoEncoder (SR-VAE)

- We consider improving the performance of OOD object detection from two perspectives:
  - One is to strengthen the discrimination ability of the object classifier for known ID objects, which is conducive to reduce the risk of misclassifying the ID objects into the OOD category
  - Another is to synthesize the virtual OOD features that significantly deviate from the distribution of the ID features, which is instrumental in boosting the performance of distinguishing OOD objects from ID objects



Structure-Enhanced Recurrent Variational AutoEncoder (SR-VAE)

- To attain these two goals, we explore exploiting Variational AutoEncoder (VAE) to separately generate the augmented ID features and virtual OOD features
- A method of SR-VAE is proposed, mainly consisting of two dedicated recurrent VAE branches



Structure Enhancement via LoG Operator

- In general, object detection involves two subtasks: object localization and classification. To this end, it is important to enhance object-related information
- We explore using the LoG operation on the extracted low-level features to strengthen the structure-relevant information

$$\mathcal{G} = \begin{bmatrix} 0 & 1 & 1 & 2 & 2 & 2 & 1 & 1 & 0 \\ 1 & 2 & 4 & 5 & 5 & 5 & 4 & 2 & 1 \\ 1 & 4 & 5 & 3 & 0 & 3 & 5 & 4 & 1 \\ 2 & 5 & 3 & -12 & -24 & -12 & 3 & 5 & 2 \\ 2 & 5 & 0 & -24 & -40 & -24 & 0 & 5 & 2 \\ 2 & 5 & 3 & -12 & -24 & -12 & 3 & 5 & 2 \\ 1 & 4 & 5 & 3 & 0 & 3 & 5 & 4 & 1 \\ 1 & 2 & 4 & 5 & 5 & 5 & 4 & 2 & 1 \\ 0 & 1 & 1 & 2 & 2 & 2 & 1 & 1 & 0 \end{bmatrix}$$

$$\mathcal{E} = F * \mathcal{G}, \qquad E = \Psi([F,\ \mathcal{E}])$$

Recurrent VAE for Improving Discrimination

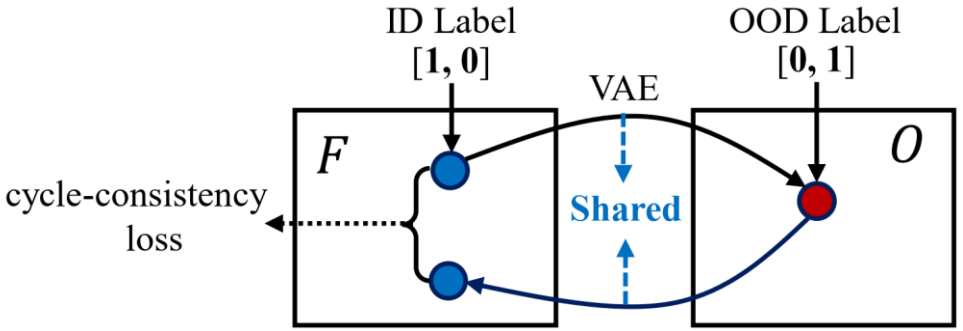
- To reduce the risk of misclassifying ID objects into the OOD category, we design a VAE module to recurrently generate diverse augmented features of the classification features, enhancing the discrimination ability

$$\mu_t = \Phi_\mu(H_{t-1}), \qquad \sigma_t = \Phi_\sigma(H_{t-1}),$$
$$h_t = \mu_t + \epsilon \cdot \exp(\sigma_t), \qquad H_t = \Theta(h_t),$$

$$\mathcal{L}_{\text{in}} = \mathcal{L}_{\text{det}} + \alpha \cdot \frac{1}{T} \sum_{t=1}^{T} \text{KL}[p(H_t|H_{t-1}), p(P_{\text{in}})],$$

Synthesizing Virtual OOD Features

- To reduce the impact of lacking OOD data, we propose a cycle-consistent conditional VAE to synthesize virtual OOD features



$$\mu_f = W_\mu * \hat{F}, \qquad \sigma_f = W_\sigma * \hat{F},$$

$$Z = \mu_f + \epsilon \cdot \exp(\sigma_f)$$

$$\mathcal{L}_{\text{cycle}} = \frac{1}{wh} \sum |\boldsymbol{F} - F|$$

Algorithm 1 SR-VAE for Unsupervised OOD-OD

Input: ID data $\{X, Y\}$, randomly initialized detector with parameter θ, weight α for the KL-loss, weight λ for the loss $\mathcal{L}_{\text{dis}}$, weight τ for the uncertainty loss $\mathcal{L}_{\text{uncertainty}}$.
Output: Object detector θ^*, and OOD classifier $\mathcal{C}$.
while *train* **do**
  Sample images from the ID dataset $\{X, Y\}$.
  Calculate the structure-enhanced map E and diverse augmented features H using Eq. (2) and (3).
  Synthesize the virtual OOD map $\boldsymbol{O}$ using Eq. (5).
  Calculate the overall training objective $\mathcal{L}$ using Eq. (4), (6), (7), and (8).
  Update the parameters θ based on Eq. (8).
end
while *eval* **do**
  Calculate the OOD uncertainty score using Eq. (9).
  Perform thresholding comparison using Eq. (9).
end

Experiments

Method (VOC)	FPR95 ↓	AUROC ↑	mAP (ID)↑
	OOD: MS-COCO / OpenImages		
MSP [14]	70.99 / 73.13	83.45 / 81.91	48.7
ODIN [28]	59.82 / 63.14	82.20 / 82.59	48.7
Mahalanobis [26]	67.73 / 65.41	81.45 / 81.48	48.7
Gram matrices [38]	62.75 / 67.42	79.88 / 77.62	48.7
Energy score [30]	56.89 / 58.69	83.69 / 82.98	48.7
Generalized ODIN [16]	59.57 / 70.28	83.12 / 79.23	48.1
CSI [42]	59.91 / 57.41	81.83 / 82.95	48.1
GAN-synthesis [25]	60.93 / 59.97	83.67 / 82.67	48.5
VOS (Baseline) [7]	47.53 / 51.33	88.70 / 85.23	48.9
SR-VAE	**42.17 / 46.26**	**90.28 / 87.89**	**49.4**
Method (BDD)	**FPR95 ↓**	**AUROC ↑**	**mAP (ID)↑**
	OOD: MS-COCO / OpenImages		
MSP [14]	80.94 / 79.04	75.87 / 77.38	31.2
ODIN [28]	62.85 / 58.92	74.44 / 76.61	31.2
Mahalanobis [26]	55.74 / 47.69	85.71 / 88.05	31.2
Gram matrices [38]	60.93 / 77.55	74.93 / 59.38	31.2
Energy score [30]	60.06 / 54.97	77.48 / 79.60	31.2
Generalized ODIN [16]	57.27 / 50.17	85.22 / 87.18	**31.8**
CSI [42]	47.10 / 37.06	84.09 / 87.99	30.6
GAN-synthesis [25]	57.03 / 50.61	78.82 / 81.25	31.4
VOS (Baseline) [7]	44.27 / 35.54	86.87 / 88.52	31.3
SR-VAE	**32.23 / 21.81**	**90.69 / 93.55**	31.5

Table 1. The performance (%) of unsupervised OOD-OD. All methods are trained based on ID data and do not use any auxiliary data. ↑ denotes larger values are better and ↓ denotes smaller values are better. We can see that our method outperforms the comparison methods significantly.

Method	AP_{novel}	AP
WSDDN [3]	19.7	19.6
Cap2Det [47]	20.3	20.1
OVR-CNN [49]	22.8	39.9
RegionCLIP [50]	26.8	47.5
Detic [51]	27.8	45.0
OCA (Baseline) [33]	36.6	49.4
OCA+Ours	**40.1**	**49.5**

Table 2. OVD results (%) on COCO. 'OCA + Ours' indicates that we directly plug our method into OCA [33].

Experiments

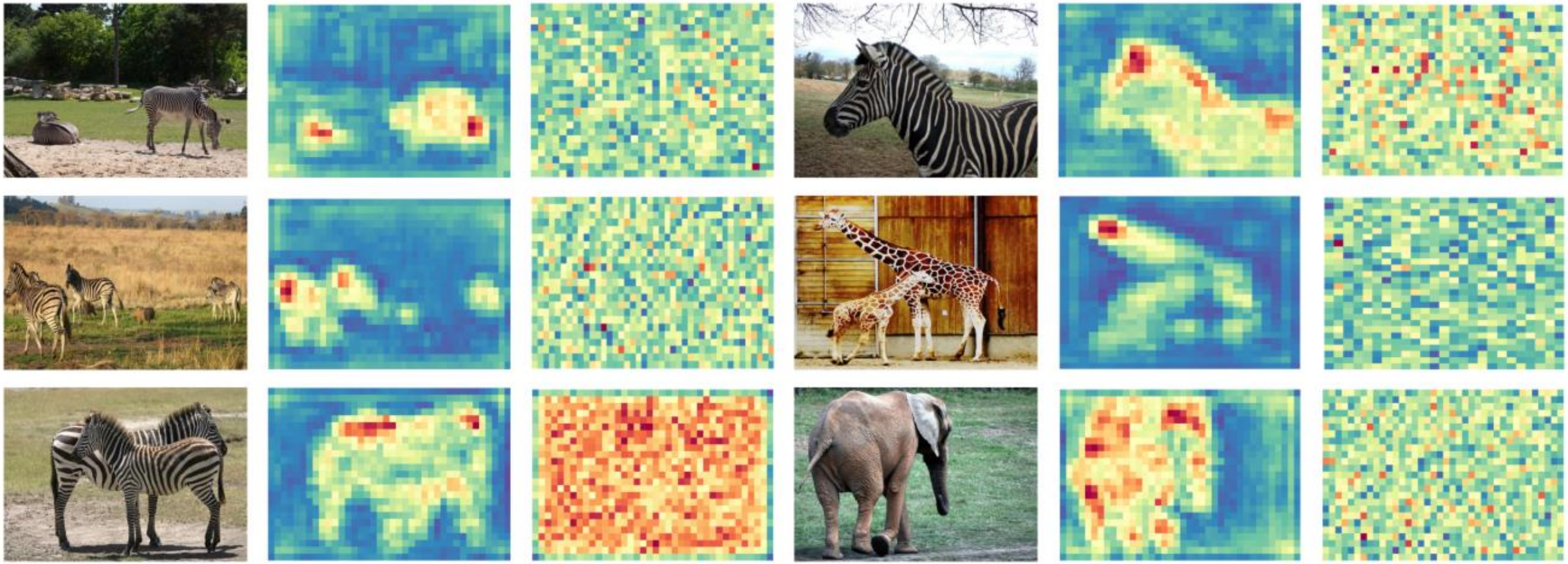
10 + 10 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
Faster ILOD (50) [32]	72.8	75.7	71.2	60.5	61.7	70.4	83.3	76.6	53.1	72.3	36.7	70.9	66.8	67.6	66.1	24.7	63.1	48.1	57.1	43.6	62.2
ORE (50) [19]	63.5	70.9	58.9	42.9	34.1	76.2	80.7	76.3	34.1	66.1	56.1	70.4	80.2	72.3	81.8	42.7	71.6	68.1	77	67.7	64.6
OW-DETR (50) [10]	61.8	69.1	67.8	45.8	47.3	78.3	78.4	78.6	36.2	71.5	57.5	75.3	76.2	77.4	79.5	40.1	66.8	66.3	75.6	64.1	65.7
ROSETTA (50) [45]	74.2	76.2	64.9	54.4	57.4	76.1	84.4	68.8	52.4	67.0	62.9	63.3	79.8	72.8	78.1	40.1	62.3	61.2	72.4	66.8	66.8
iOD (50) [22]	76.0	74.6	67.5	55.9	57.6	75.1	85.4	77.0	43.7	70.8	60.1	66.4	76.0	72.6	74.6	39.7	64.0	60.2	68.5	60.5	66.3
iOD + Ours (50)	75.9	75.2	68.8	55.3	55.5	77.7	85.6	79.3	49.4	78.2	61.0	75.3	81.4	74.5	79.3	43.8	72.5	67.0	70.2	65.7	**69.6**
iOD (75) [22]	39.0	36.5	28.4	19.4	24.2	47.2	56.7	41.0	19.1	48.0	21.1	32.1	43.0	36.3	40.0	14.8	40.1	36.5	37.3	45.3	35.3
iOD + Ours (75)	43.6	41.0	31.3	24.9	29.8	55.4	60.8	44.1	22.4	46.7	29.5	32.3	35.6	38.3	35.7	15.1	46.9	34.6	37.9	46.7	**37.6**

15 + 5 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
Faster ILOD (50) [32]	66.5	78.1	71.8	54.6	61.4	68.4	82.6	82.7	52.1	74.3	63.1	78.6	80.5	78.4	80.4	36.7	61.7	59.3	67.9	59.1	67.9
ORE (50) [19]	75.4	81.0	67.1	51.9	55.7	77.2	85.6	81.7	46.1	76.2	55.4	76.7	86.2	78.5	82.1	32.8	63.6	54.7	77.7	64.6	68.5
OW-DETR (50) [10]	77.1	76.5	69.2	51.3	61.3	79.8	84.2	81.0	49.7	79.6	58.1	79.0	83.1	67.8	85.4	33.2	65.1	62.0	73.9	65.0	69.4
ROSETTA (50) [45]	76.5	77.5	65.1	56.0	60.0	78.3	85.5	78.7	49.5	68.2	67.4	71.2	83.9	75.7	82.0	43.0	60.6	64.1	72.8	67.4	69.2
iOD (50) [22]	78.4	79.7	66.9	54.8	56.2	77.7	84.6	79.1	47.7	75.0	61.8	74.7	81.6	77.5	80.2	37.8	58.0	54.6	73.0	56.1	67.8
iOD + Ours (50)	78.3	80.3	70.5	51.6	60.2	79.4	85.9	76.2	52.5	79.4	65.2	81.8	83.7	76.1	77.9	41.1	62.8	63.8	72.6	67.9	**70.4**
iOD (75) [22]	40.7	40.9	28.7	19.1	23.8	61.6	56.1	38.8	23.6	47.5	18.7	40.1	40.2	41.5	39.8	9.1	40.6	32.4	41.9	47.6	36.6
iOD + Ours (75)	44.4	44.5	36.5	21.2	27.6	55.5	63.7	39.8	24.9	50.3	27.2	41.6	47.9	43.9	41.4	11.3	39.1	38.6	43.1	48.5	**39.5**

19 + 1 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
Faster ILOD (50) [32]	64.2	74.7	73.2	55.5	53.7	70.8	82.9	82.6	51.6	79.7	58.7	78.8	81.8	75.3	77.4	43.1	73.8	61.7	69.8	61.1	68.6
ORE (50) [19]	67.3	76.8	60.0	48.4	58.8	81.1	86.5	75.8	41.5	79.6	54.6	72.8	85.9	81.7	82.4	44.8	75.8	68.2	75.7	60.1	68.9
OW-DETR (50) [10]	70.5	77.2	73.8	54.0	55.6	79.0	80.8	80.6	43.2	80.4	53.5	77.5	89.5	82.0	74.7	43.3	71.9	66.6	79.4	62.0	70.2
ROSETTA (50) [45]	75.3	77.9	65.3	56.2	55.3	79.6	84.6	72.9	49.2	73.7	68.3	71.0	78.9	77.7	80.7	44.0	69.6	68.5	76.1	68.3	69.6
iOD (50) [22]	78.2	77.5	69.4	55.0	56.0	78.4	84.2	79.2	46.6	79.0	63.2	78.5	82.7	79.1	79.9	44.1	73.2	66.3	76.4	57.6	70.2
iOD + Ours (50)	76.6	83.5	74.7	57.0	58.0	77.0	85.6	82.5	51.5	82.7	61.4	81.6	82.9	79.8	77.6	47.4	74.7	68.4	74.1	59.0	**71.8**
iOD (75) [22]	35.9	44.7	31.6	22.4	26.9	52.0	56.5	38.7	21.6	48.4	21.2	35.9	37.9	30.7	38.7	17.2	38.5	34.2	40.7	46.6	36.0
iOD + Ours (75)	36.4	45.1	36.1	18.0	28.9	53.2	62.2	38.5	25.3	55.1	27.4	46.8	45.9	42.9	40.3	20.9	50.8	37.0	44.4	47.1	**40.1**

Table 3. Performance (%) analysis of class-incremental object detection. 'iOD + Ours' indicates that our method is plugged into iOD [22]. Here, '50' and '75' separately represent that the mAP metric is calculated when the IOU threshold is set to 0.5 and 0.75.

Experiments



(a) Input Image (b) Structure Map (c) Virtual OOD Map (d) Input Image (e) Structure Map (f) Virtual OOD Map

Thanks for Your Listening!