

# Latent Outlier Exposure For Anomaly Detection With Contaminated Data

Chen Qiu<sup>\*1,2</sup>, Aodong Li<sup>\*3</sup>, Marius Kloft<sup>2</sup>, Maja Rudolph<sup>1</sup>, Stephan Mandt<sup>3</sup>

1: Bosch Center for AI

2: Technical University Kaiserslautern, Germany

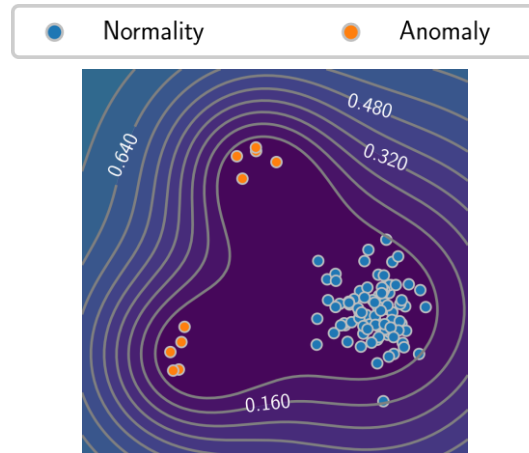
3: University of California, Irvine, USA

\* Equal Contribution

# Deep Anomaly Detection with Contaminated Data

## Motivation

- ▶ A common assumption in unsupervised anomaly detection is that **clean** training data is available.
- ▶ What if the training data is not clean and already contains some anomalies?
  - ▶ A dataset of financial transactions could contain unnoticed fraudulent activities



Blind update  
on unlabeled dataset



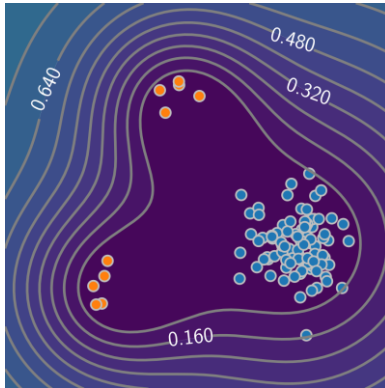
A clear performance degradation

# Deep Anomaly Detection with Contaminated Data

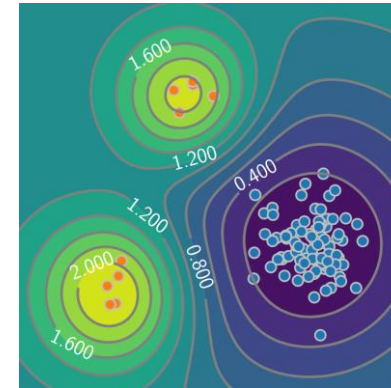
## Motivation

- ▶ When the contaminated data is well labeled, the model learns the decision boundary very well.
- ▶ Unexploited information is hidden in the contaminated data.

● Normality      ● Anomaly



Blind update  
on unlabeled dataset



Supervised update  
on labeled dataset

# Deep Anomaly Detection with Contaminated Data

## Contribution & Problem setup

- Our contribution: **Latent Outlier Exposure (LOE)**, a new general strategy for training deep anomaly detection models with contaminated data.
- Problem setup: we consider a set of samples  $x_i$  that contains many normal samples and a few anomalies.



# Latent Outlier Exposure

## Proposed loss

- Consider a joint loss function with two loss terms (given the binary anomaly label  $y_i$ ):

$$\mathcal{L}(\theta, \mathbf{y}) = \sum_{i=1}^N (1 - y_i) \mathcal{L}_n^\theta(\mathbf{x}_i) + y_i \mathcal{L}_a^\theta(\mathbf{x}_i).$$

- $\mathcal{L}_n^\theta(\mathbf{x})$  : a normal loss that is designed to be minimized over normal data.
- $\mathcal{L}_a^\theta(\mathbf{x})$  : an abnormal loss that is designed to have the opposite effect.
- E.g., for deep SVDD,  $\mathcal{L}_n^\theta(\mathbf{x}) = \|f_\theta(\mathbf{x}) - \mathbf{c}\|^2$  and  $\mathcal{L}_a^\theta(\mathbf{x}) = 1/\|f_\theta(\mathbf{x}) - \mathbf{c}\|^2$ .

# Latent Outlier Exposure Optimization

- ▶ Problem: the anomaly labels  $y$  are not observed in practice.
- ▶ Key idea: infer the anomaly labels during training
- ▶ Constrained minimization problem:

$$\min_{\theta} \min_{y \in \mathcal{Y}} \mathcal{L}(\theta, y) \quad \text{s.t.} \quad \mathcal{Y} = \{y \in \{0, 1\}^N : \sum_{i=1}^N y_i = \alpha N\}$$

where  $\alpha$  is an assumed contamination rate (hyperparameter), and  $N$  is the number of samples.

# Latent Outlier Exposure Optimization

Block coordinate update :

- ▶ To update  $\theta$ , we fix  $y$  and minimize  $\mathcal{L}(\theta, y)$  over  $\theta$ .
- ▶ To infer  $y$  given  $\theta$ , we minimize  $\mathcal{L}(\theta, y)$  subject to the constraint.
  - ▶ Rank samples according to  $S_i^{train} = \mathcal{L}_n^\theta(\mathbf{x}_i) - \mathcal{L}_a^\theta(\mathbf{x}_i)$  and assign 1 to the top  $\alpha$  fraction, and 0 to the rest.

# Latent Outlier Exposure

## Model extension & Anomaly score

- Soft latent outlier exposure:

$$\mathcal{Y}' = \{\mathbf{y} \in \{0, 0.5\}^N : \sum_{i=1}^N y_i = 0.5\alpha N\}.$$

- Avoid being overconfident in assigning labels
- Treat the identified anomalies as uncertain

- Anomaly score at the test time:

$$S_i^{test} = \mathcal{L}_n^\theta(\mathbf{x}_i).$$

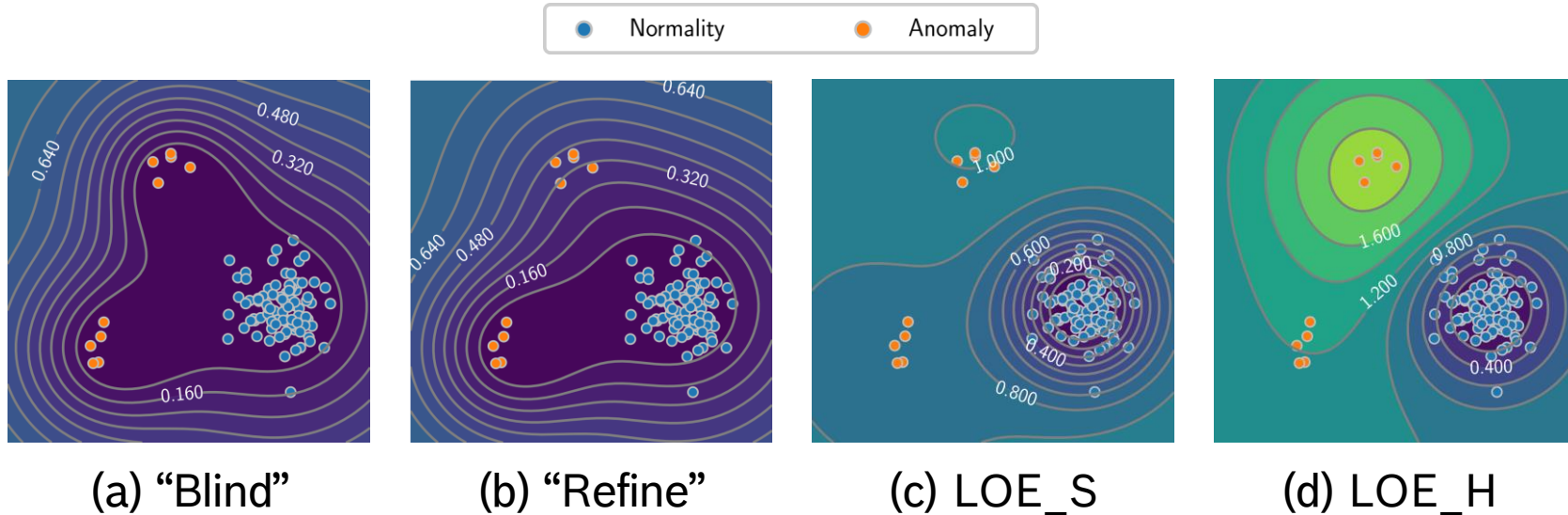
- Avoid the overfitting to anomalies encountered during training



# Latent Outlier Exposure

## Toy experiment

- Soft LOE and hard LOE learn the boundary of normal and abnormal samples successfully, while the “Blind” and “Refine” baselines do not.



# Latent Outlier Exposure

## Experiment setup

- Datasets: image data, video data, and tabular data
- Corrupt the training set: we mix a fraction of  $\alpha_0$  abnormal samples into the normal training set.

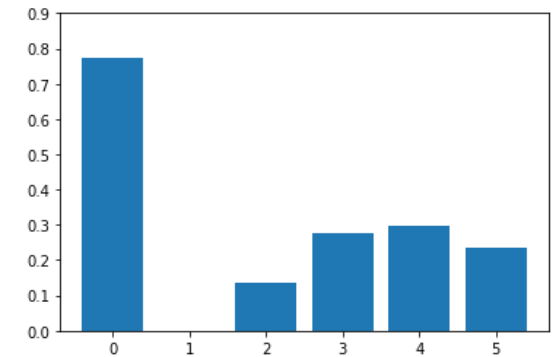
Image data



Video data



Tabular data



# Latent Outlier Exposure

## Quantitative results

- ▶ LOE outperforms the baselines on image data, video data, and tabular data when combining with various deep anomaly detection models.

Table 1. AUC (%) with standard deviation for anomaly detection on CIFAR-10 and F-MNIST. For all experiments, we set the contamination ratio as 10%. LOE mitigates the performance drop when NTL and MHRot trained on the contaminated datasets.

		CIFAR-10	F-MNIST
NTL	Blind	91.3±0.1 (-4.4)	85.0±0.2 (-9.7)
	Refine	93.5±0.1 (-2.2)	89.1±0.2 (-5.6)
	LOE <sub>H</sub> (ours)	<b>94.9±0.2 (-0.8)</b>	<b>92.9±0.7 (-1.8)</b>
	LOE <sub>S</sub> (ours)	<b>94.9±0.1 (-0.8)</b>	92.5±0.1 (-2.2)
MHRot	Blind	84.0±0.5 (-4.2)	88.8±0.1 (-4.9)
	Refine	84.4±0.1 (-3.8)	89.6±0.2 (-4.1)
	LOE <sub>H</sub> (ours)	<b>86.4±0.5 (-1.8)</b>	<b>91.4±0.2 (-2.3)</b>
	LOE <sub>S</sub> (ours)	86.3±0.2 (-1.9)	91.2±0.4 (-2.5)

Table 4. AUC (%) for different contamination ratios for a video frame anomaly detection benchmark proposed in (Pang et al., 2020). LOE<sub>S</sub> (proposed) achieves state-of-the-art performance.

Method	Contamination Ratio		
	10%	20%	30%*
(Tudor Ionescu et al., 2017)	-	-	68.4
(Liu et al., 2018)	-	-	69.0
(Del Giorno et al., 2016)	-	-	59.6
(Sugiyama & Borgwardt, 2013)	55.0	56.0	56.3
(Pang et al., 2020)	68.0	70.0	<b>71.7</b>
Blind	85.2±1.0	76.0±2.7	66.6±2.6
Refine	82.7±1.5	74.9±2.4	69.3±0.7
LOE <sub>H</sub> (ours)	82.3±1.6	59.6±3.8	56.8±9.5
LOE <sub>S</sub> (ours)	<b>86.8±1.2</b>	<b>79.2±1.3</b>	<b>71.5±2.4</b>

\*Default setup in (Pang et al., 2020), corresponding to  $\alpha_0 \approx 30\%$ .

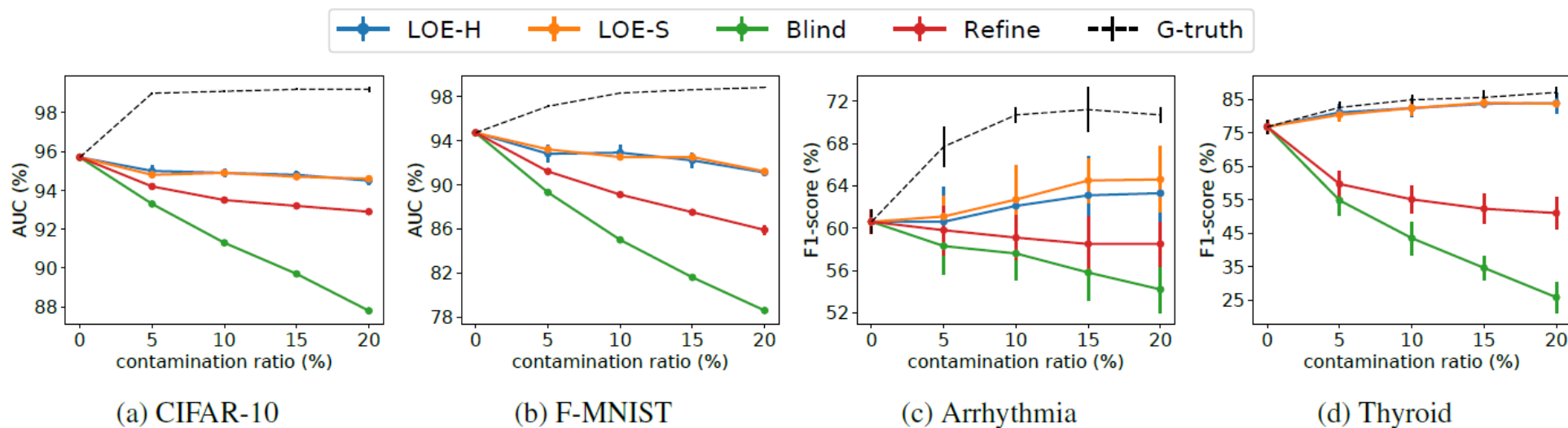
Table 3. F1-score (%) for anomaly detection on 30 tabular datasets studied in (Anonymous, 2022). We set  $\alpha_0 = \alpha = 10\%$  in all experiments. LOE (proposed) outperforms the “Blind” and “Refine” consistently. (See Tables 5 and 6 for more details, including AUCs.)

	NTL				ICL			
	Blind	Refine	LOE <sub>H</sub> (ours)	LOE <sub>S</sub> (ours)	Blind	Refine	LOE <sub>H</sub> (ours)	LOE <sub>S</sub> (ours)
abalone	37.9±13.4	55.2±15.9	42.8±26.9	<b>59.3±12.0</b>	50.9±1.5	<b>54.3±2.9</b>	53.4±5.2	51.7±2.4
anthyroid	29.7±3.5	42.7±7.1	47.7±11.4	<b>50.3±4.5</b>	29.1±2.2	38.5±2.1	<b>48.7±7.6</b>	43.0±8.8
arrhythmia	57.6±2.5	59.1±2.1	62.1±2.8	<b>62.7±3.3</b>	53.9±0.7	60.9±2.2	62.4±1.8	<b>63.6±2.1</b>
breastw	84.0±1.8	93.1±0.9	<b>95.6±0.4</b>	95.3±0.4	92.6±1.1	93.4±1.0	<b>96.0±0.6</b>	95.7±0.6
cardio	21.8±4.9	45.2±7.9	<b>73.0±7.9</b>	57.8±5.5	50.2±4.5	56.2±3.4	<b>71.1±3.2</b>	62.2±2.7
ecoli	0.0±0.0	88.9±14.1	<b>100.0±0.0</b>	<b>100.0±0.0</b>	17.8±15.1	46.7±25.7	<b>75.6±4.4</b>	<b>75.6±4.4</b>
forest cover	20.4±4.0	56.2±4.9	61.1±34.9	<b>67.6±30.6</b>	9.2±4.5	8.0±3.6	6.8±3.6	<b>11.1±2.1</b>
glass	11.1±7.0	15.6±5.4	17.8±5.4	<b>20.0±8.3</b>	8.9±4.4	<b>11.1±0.0</b>	<b>11.1±7.0</b>	8.9±8.3
ionosphere	89.0±1.5	91.0±2.0	91.0±1.7	<b>91.3±2.2</b>	86.5±1.1	85.9±2.3	85.7±2.8	<b>88.6±0.6</b>
kdd	95.9±0.0	96.0±1.1	98.1±0.4	<b>98.4±0.1</b>	99.3±0.1	99.4±0.1	<b>99.5±0.0</b>	99.4±0.0
kddrev	98.4±0.1	98.4±0.2	89.1±1.7	<b>98.6±0.0</b>	97.9±0.5	98.4±0.4	<b>98.8±0.1</b>	98.2±0.4
letter	36.4±3.6	44.4±3.1	25.4±10.0	<b>45.6±10.6</b>	43.0±2.5	51.2±3.7	<b>54.4±5.6</b>	47.2±4.9
lympho	53.3±12.5	60.0±8.2	60.0±13.3	<b>73.3±22.6</b>	43.3±8.2	60.0±8.2	80.0±12.5	<b>83.3±10.5</b>
mammogra.	5.5±2.8	2.6±1.7	3.3±1.6	<b>13.5±3.8</b>	8.8±1.9	11.4±1.9	34.0±20.2	<b>42.8±17.6</b>
mnist tabular	78.6±0.5	<b>80.3±1.1</b>	71.8±1.8	76.3±2.1	72.1±1.0	80.7±0.7	<b>86.0±0.4</b>	79.2±0.9
mulcross	45.5±9.6	<b>58.2±3.5</b>	<b>58.2±6.2</b>	50.1±8.9	70.4±13.4	94.4±6.3	<b>100.0±0.0</b>	99.9±0.1
musk	21.0±3.3	98.8±0.4	<b>100.0±0.0</b>	<b>100.0±0.0</b>	6.2±3.0	<b>100.0±0.0</b>	<b>100.0±0.0</b>	<b>100.0±0.0</b>
optdigits	0.2±0.3	1.5±0.3	41.7±45.9	<b>59.1±48.2</b>	0.8±0.5	<b>1.3±1.1</b>	1.2±1.0	0.9±0.5
pendigits	5.0±2.5	32.6±10.0	79.4±4.7	<b>81.9±4.3</b>	10.3±4.6	30.1±8.5	80.3±6.1	<b>88.6±2.2</b>
pima	60.3±2.6	61.0±1.9	<b>61.3±2.4</b>	61.0±0.9	58.1±2.9	59.3±1.4	<b>63.0±1.0</b>	60.1±1.4
satellite	73.6±0.4	74.1±0.3	<b>74.8±0.4</b>	74.7±0.1	72.7±1.3	72.7±0.6	<b>73.6±0.2</b>	73.2±0.6
satimage	26.8±1.5	86.8±4.0	90.7±1.1	<b>91.0±0.7</b>	7.3±0.6	85.1±1.4	91.3±1.1	<b>91.5±0.9</b>
seismic	11.9±1.8	11.5±1.0	<b>18.1±0.7</b>	17.1±0.6	14.9±1.4	17.3±2.1	23.6±2.8	<b>24.2±1.4</b>
shuttle	97.0±0.3	97.0±0.2	<b>97.1±0.2</b>	97.0±0.2	96.6±0.2	96.7±0.1	96.9±0.1	<b>97.0±0.2</b>
speech	6.9±1.2	8.2±2.1	43.3±5.6	<b>50.8±2.5</b>	0.3±0.7	1.6±1.0	<b>2.0±0.7</b>	0.7±0.8
thyroid	43.4±5.5	55.1±4.2	<b>82.4±2.7</b>	<b>82.4±2.3</b>	45.8±7.3	71.6±2.4	<b>83.2±2.9</b>	80.9±2.5
vertebral	22.0±4.5	21.3±4.5	22.7±11.0	<b>25.3±4.0</b>	8.9±3.1	8.9±4.2	7.8±4.2	<b>10.0±2.7</b>
vowels	36.0±1.8	50.4±8.8	<b>62.8±9.5</b>	48.4±6.6	42.1±9.0	60.4±7.9	<b>81.6±2.9</b>	74.4±8.0
wbc	25.7±12.3	45.7±15.5	<b>76.2±6.0</b>	69.5±3.8	50.5±5.7	50.5±2.3	<b>61.0±4.7</b>	<b>61.0±1.9</b>
wine	24.0±18.5	66.0±12.0	90.0±0.0	<b>92.0±4.0</b>	4.0±4.9	10.0±8.9	98.0±4.0	<b>100.0±0.0</b>

# Latent Outlier Exposure

## Experiment results on various contamination ratios

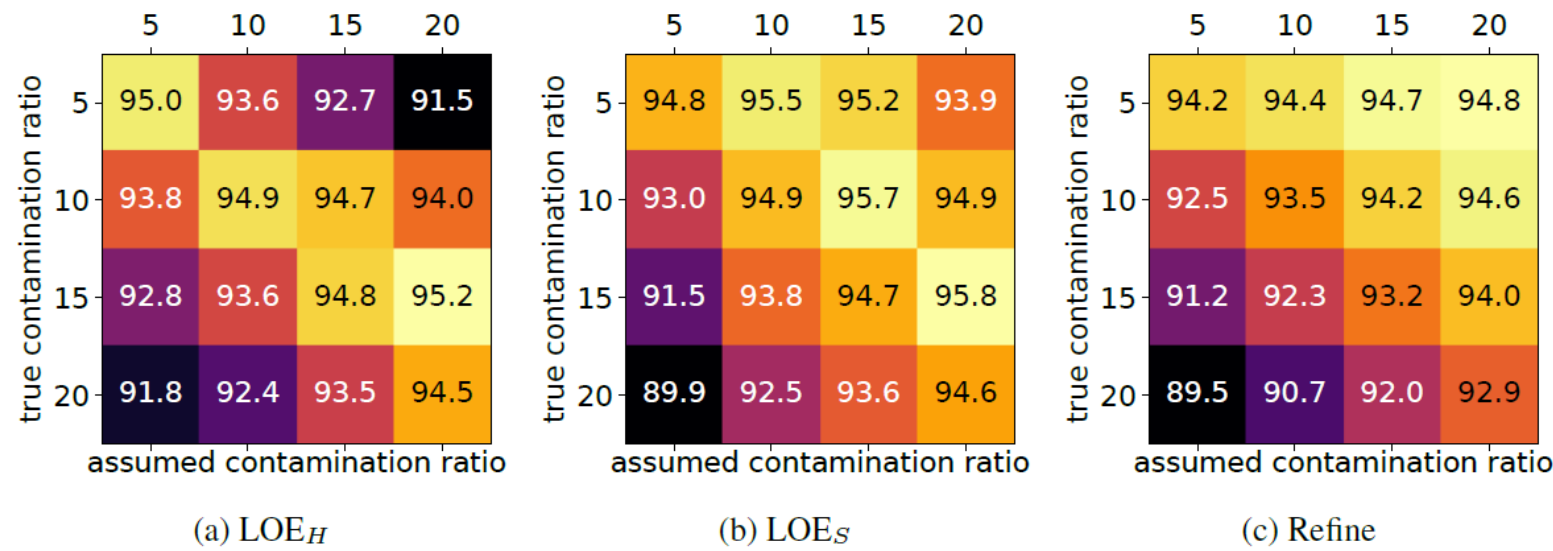
- LOE shows considerable robustness to different contamination ratios.



# Latent Outlier Exposure

## Sensitivity study on the misspecified contamination ratio

- LOE shows considerable robustness against misspecified contamination ratios.



# THANK YOU

Paper Link



Code Link



Correspondence: [Chen.Qiu@de.bosch.com](mailto:Chen.Qiu@de.bosch.com)