DLSR

Diffusion-based Layer-wise Semantic Reconstruction for Unsupervised Out-of-Distribution Detection

21.5%
AUROC increase
Compared to DDPM
(Pixel level)

FASTER
100x
faster
than pixel level

EASIER
To Follow

Ying Yang[1], De Cheng[1†], Chaowei Fang[1†], Yubiao Wang[1],Changzhe Jiao[1], Lechao Cheng[2], Nannan Wang[1], Xinbo Gao[3]

NeurIPS 2024

DLSR

Diffusion-based Layer-wise Semantic Reconstruction for Unsupervised Out-of-Distribution Detection

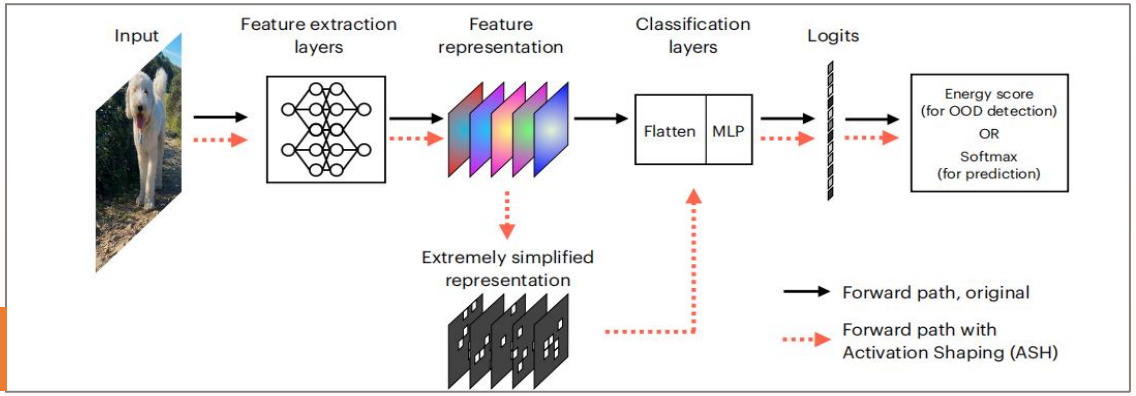
Ying Yang[1], De Cheng[1†], Chaowei Fang[1†], Yubiao Wang[1], Changzhe Jiao[1], Lechao Cheng[2], Nannan Wang[1], Xinbo Gao[3]

Motivation:

Improving the reconstruction power of the generative model, while keeping compact representation of the ID data

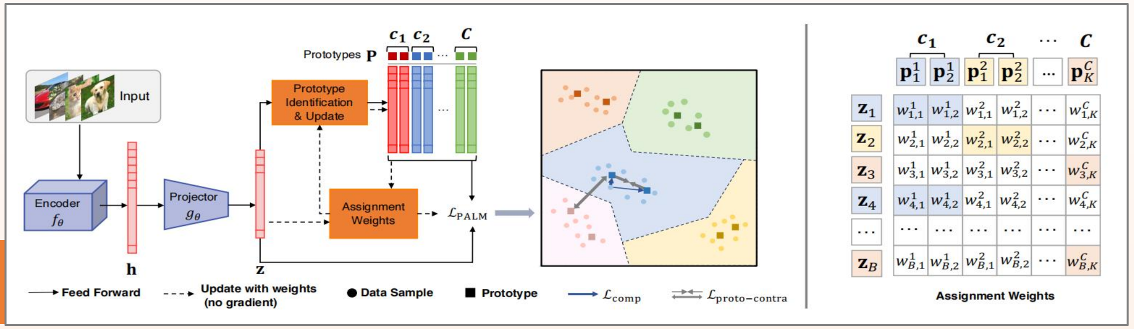
Related Works

Classification Based



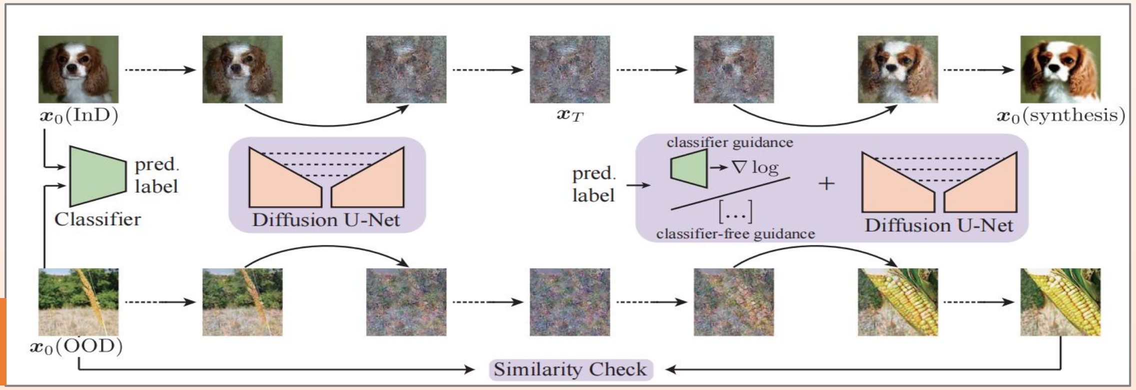
Encounter issues with **assigning high softmax probability** to OOD samples

Distance Based



Fail to **capture sample distribution** accurately.

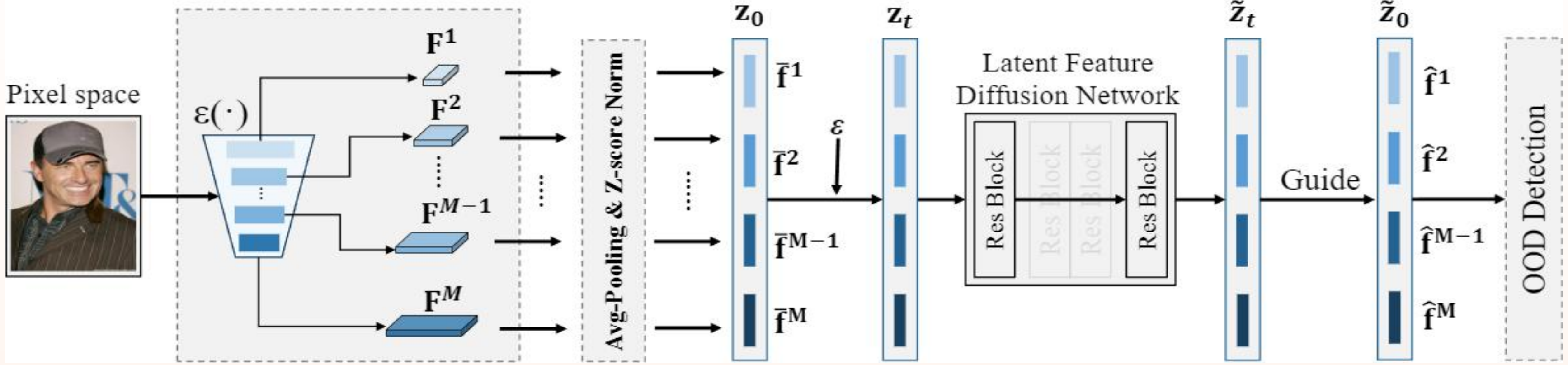
Pixel-Gen Based



Significantly **high training and inference time costs.**

Proposed Method: DLSR--(Feature-Gen Based)

LFDN



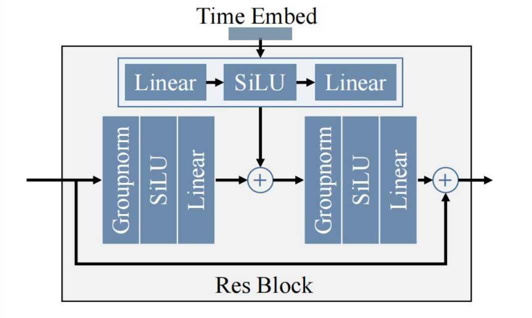
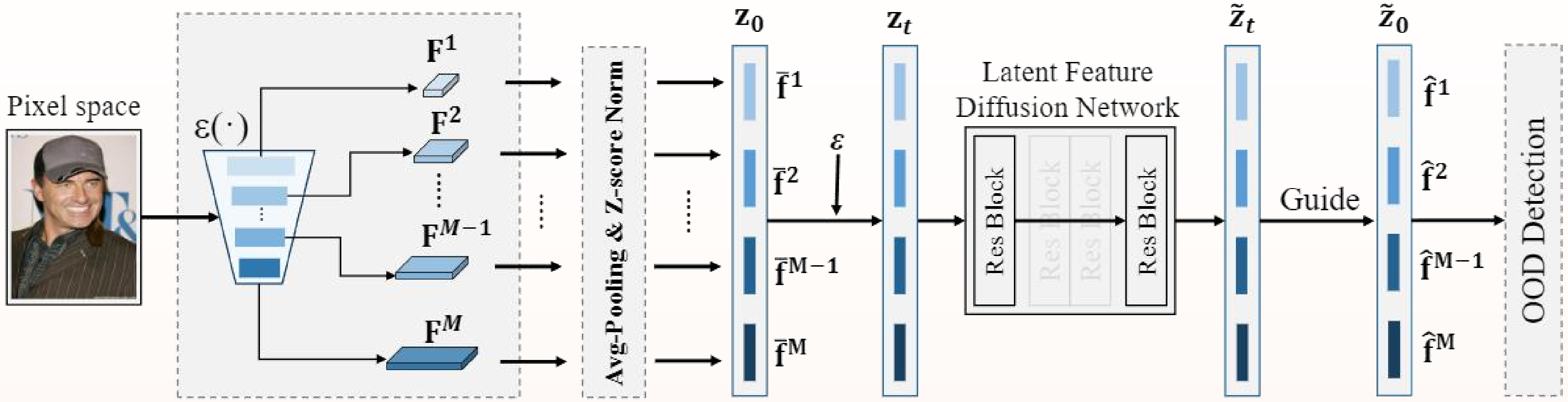


Figure 2: Residual Block Structure in LFDN.

- We propose a diffusion-based layer-wise semantic reconstruction framework to tackle OOD detection, based on multi-layer semantic feature distortion and reconstruction. Meanwhile, We are the first to successfully incorporate generative modeling of features within the framework of OOD detection in image classification tasks.
- The layer-wise semantic feature reconstruction encourages restricting the in-distribution latent features to be more compactly distributed within a certain space, enabling better reconstruction of ID samples while limiting the reconstruction of OOD samples.
- Extensive experiments on multiple benchmarks across various datasets show that our method achieves state-of-the-art detection accuracy and speed.



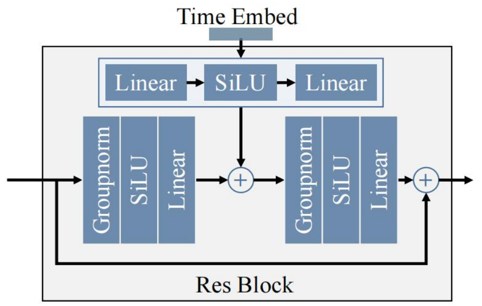


Figure 2: Residual Block Structure in LFDN.

1. Multi-layer Semantic Feature Extraction

$$\mathbf{F}^m \in \mathbb{R}^{c_m \times w_m \times h_m}, m \in \{1, ..., M\}$$

$$\bar{\mathbf{f}}^m = \frac{\mathbf{f}^m - \mu_{\mathbf{f}^m}}{\sqrt{\mathrm{Var}(\mathbf{f}^m)+\delta}} \qquad c = \sum_{m=1}^{M} c_m$$

$$\mathbf{z}_0 = \mathcal{H}(\mathbf{x}) = [\bar{\mathbf{f}}^1, \ldots, \bar{\mathbf{f}}^m, \ldots, \bar{\mathbf{f}}^M] \in \mathbb{R}^c$$

2. Diffusion-based Feature Distortion and Reconstruction

$$\mathbf{z}_t = \mathrm{ennoise}(\mathbf{z}_0, t) = \sqrt{\bar{\alpha}_t} \times \mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t} \times \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}^c, \mathbf{I}^{c\times c})$$

$$\tilde{\boldsymbol{\epsilon}}_t = \frac{\mathbf{z}_t - \sqrt{\bar{\alpha}_t} \times \tilde{\mathbf{z}}_t}{\sqrt{1-\bar{\alpha}_t}}, \qquad \text{Loss: } L = \frac{1}{N}\sum_{\mathbf{x}\in\mathbb{D}} \|\mathbf{z}_0 - \mathrm{LFDN}(\mathbf{z}_t, t)\|_2^2$$

$$\tilde{\mathbf{z}}_{t'} = \sqrt{\bar{\alpha}_{t'}}\left(\frac{\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t} \times \tilde{\boldsymbol{\epsilon}}_t}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t'} - \sigma_t^2} \times \tilde{\boldsymbol{\epsilon}}_t + \sigma_t^2 \boldsymbol{\epsilon},$$

3. Mean Squared Eror (MSE), Likelihood Regrat metric, Muiti-layer Semantic Feature Similarity (MFsim):

$$\mathrm{LR} = \mathrm{MSE}_{\text{initial}} - \mathrm{MSE}_{\text{final}} \qquad \mathrm{Sim}(\bar{\mathbf{f}}^m, \tilde{\mathbf{f}}^m) = \frac{\bar{\mathbf{f}}^m \cdot \tilde{\mathbf{f}}^m}{\|\bar{\mathbf{f}}^m\| \cdot \|\tilde{\mathbf{f}}^m\|}$$

Experimental & Visualization Results

Dataset		Pixel-Generative-Base				Feature-Generative-Base			
ID	OOD	GLOW	PixelCNN++	VAE	DDPM	AutoEncoder	our(+MSE)	ours(+LR)	ours(+MFsim)
CIFRA10	SVHN	88.3	73.7	95.9	97.3	57.7	97.3±0.0	98.2±0.0	**98.9±0.1**
	LSUN	21.3	64.0	40.3	68.2	81.5	97.6±0.1	97.8±0.1	**99.8±0.1**
	MNIST	85.8	96.7	**99.9**	83.2	95.8	99.4±0.0	98.9±0.1	**99.9±0.0**
	FMNIST	71.2	90.7	99.1	84.3	79.6	99.0±0.0	98.8±0.0	**99.9±0.0**
	KMNIST	38.0	82.6	**99.9**	89.7	90.5	99.5±0.0	99.1±0.0	**99.9±0.0**
	Omniglot	95.5	98.9	99.6	35.9	81.5	99.1±0.1	97.1±0.1	**99.9±0.0**
	NotMNIST	53.9	82.6	99.4	88.7	81.6	99.8±0.1	99.5±0.0	**99.9±0.0**
	average	64.9	84.2	90.6	78.2	81.2	98.8±0.1	98.5±0.1	**99.7±0.1**
Time	Num img/s (↑)	38.6	19.3	0.7	11.4	1224.2	999.3	273.6	999.3

Dataset		Pixel-Generative-Based		Feature-Generative-Based			
ID	OOD	VAE	DDPM	AutoEncoder	ours(+MSE)	ours(+LR)	ours(+MFsim)
CelebA	SUN	95.89	83.41	32.90	**99.98±0.01**	97.15±0.02	**99.98±0.01**
	iNaturalist	95.52	82.38	41.56	**100+0.00**	99.96±0.01	99.99±0.00
	Textures	91.73	78.33	56.33	99.93±0.02	98.51±0.02	**99.96±0.01**
	Places365	97.58	76.25	35.90	99.96±0.01	97.47±0.03	**99.98±0.00**
	average	95.18	80.09	41.67	99.97±0.01	98.27±0.02	**99.98±0.01**
Time	Num img/s (↑)	18.7	10.2	1357.6	1033.8	290.4	1033.8

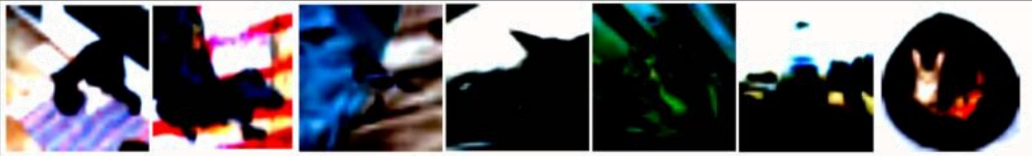


Figure 10: Examples of ID Samples Misclassified as OOD (Lacking Semantic Information).

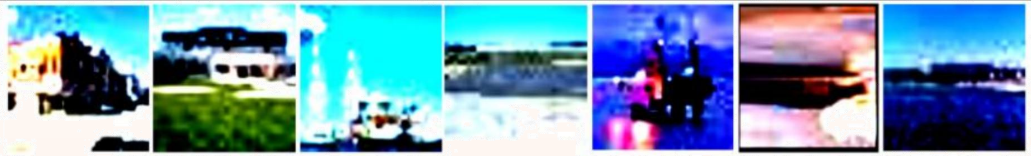


Figure 11: Examples of OOD Samples Misclassified as ID (Similar to ID Sample Categories).

Figure 12: Examples of OOD Samples Misclassified as ID (Similar to ID Sample Colors).

Our method **achieves 20.4% higher AUROC** than DDPM. This indirectly indicates that performing OOD detection at the pixel level is much worse than performing OOD detection at the feature level.

Compare
Classification-based and Distance-based Methods

ID	Based	Method	Num img/s (↑)	OOD						average
				SVHN	LSUN-c	LSUN-r	iSUN	Textures	Places365	
CIFAR10	Classifier-based	MSP	1060.5	94.53	96.37	91.80	92.23	95.93	97.59	94.74
		EBO	1060.5	96.79	97.34	94.42	94.64	96.30	98.34	96.31
		DICE	1066.3	98.53	99.03	94.49	95.25	97.68	99.63	97.44
		ASH-S	1047.6	98.01	98.23	93.17	94.13	97.01	98.48	96.51
	Distance-based	SimCLR+Mahalanobis	674.8	97.80	73.61	69.28	88.63	76.47	67.42	78.87
		SimCLR+KNN	919.8	92.40	92.05	89.81	90.14	97.24	94.36	92.67
	Generative-based	ours(+MSE)	960.6	97.31±0.02	97.59±0.01	93.93±0.01	92.78±0.01	**100±0.00**	99.96±0.00	96.93±0.01
		ours(+LR)	360.2	98.22±0.02	97.84±0.02	95.37±0.01	94.31±0.02	**100±0.00**	99.91±0.01	97.61±0.02
		ours(+MFsim)	960.6	**98.89±0.01**	**99.83±0.02**	**98.83±0.01**	**98.52±0.02**	**100±0.00**	**100±0.00**	**99.34±0.01**
CIFAR100	Classifier-based	MSP	1060.5	77.56	84.03	72.09	71.52	90.02	89.00	80.70
		EBO	1060.5	76.51	81.59	78.92	76.38	79.38	83.07	79.31
		DICE	1066.3	86.93	88.54	71.97	71.29	92.83	90.78	83.72
		ASH-S	1047.6	92.11	90.03	63.30	65.12	95.25	92.99	83.13
	Distance-based	SimCLR+Mahalanobis	674.8	56.24	52.23	61.34	73.53	71.92	51.98	61.21
		SimCLR+KNN	919.8	54.37	51.49	83.80	77.21	53.31	54.43	62.44
	Generative-based	ours(+MSE)	960.6	83.93±0.01	86.86±0.01	75.38±0.01	71.99±0.02	99.99±0.00	99.97±0.01	86.35±0.01
		ours(+LR)	360.2	88.84±0.01	87.60±0.02	80.96±0.01	77.71±0.02	99.98±0.01	99.92±0.02	89.17±0.01
		ours(+MFsim)	960.6	**93.90±0.01**	**99.14±0.01**	**95.74±0.01**	**94.40±0.01**	**100±0.00**	**100±0.00**	**97.20±0.01**

Specifically, for CIFAR-100 as the in-distribution dataset, our method integrated with MFsim achieves **an average AUROC of 13.84% higher** than the classification-based method DICE. Moreover, unlike classification-based methods, our approach does not require labeled data.

Ablation Study:

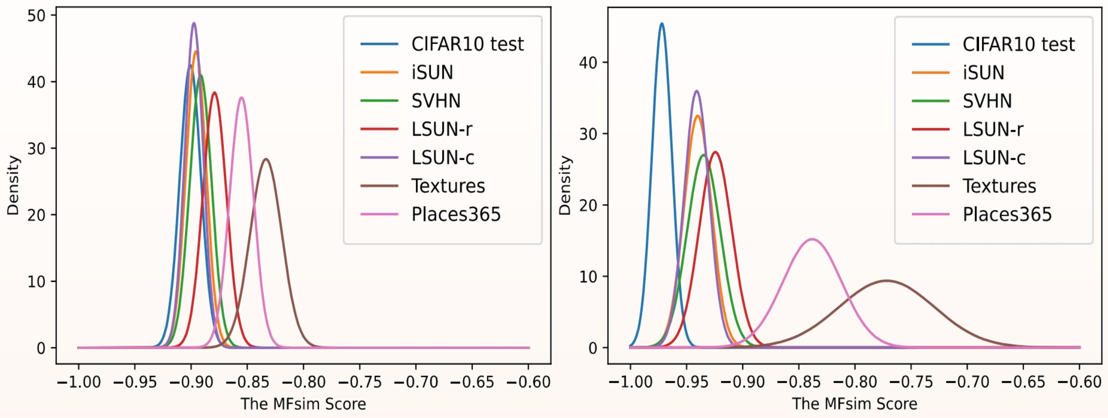


Figure 3:
The MFsim score distributions of **the first epoch (left)** and **the last epoch (right)**

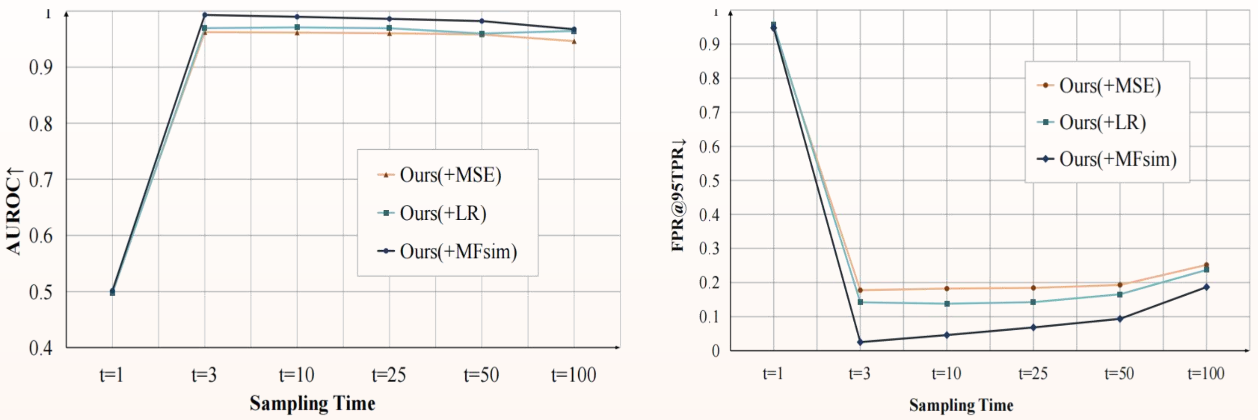


Figure4:
The average AUROC and FPR95 for the three metrics are evaluated at **different sampling time steps**.

Table 4: Changes in Average **AUROC Across Six Datasets** listed in Table 3 for CIFAR100 as ID.

Metrics	**MSE**		**LR**		**MFsim**	
Linear	Linear=720	Linear=1440	Linear=720	Linear=1440	Linear=720	Linear=1440
Average	83.35	86.35	84.05	89.17	96.43	97.20
Number of Blocks	Number=8	Number=16	Number=8	Number=16	Number=8	Number=16
Average	85.26	86.35	87.32	89.17	97.13	97.20

Diffusion-based Layer-wise Semantic Reconstruction for Unsupervised Out-of-Distribution Detection

Ying Yang[1], De Cheng[1†], Chaowei Fang[1†], Yubiao Wang[1], Changzhe Jiao[1], Lechao Cheng[2], Nannan Wang[1], Xinbo Gao[3]

NEURAL INFORMATION PROCESSING SYSTEMS