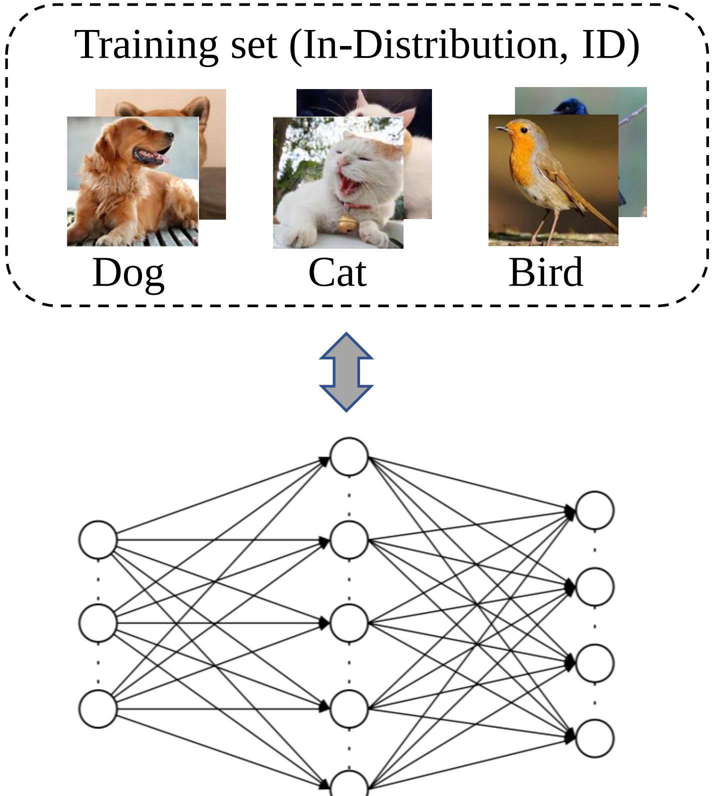
Simplification Is All You Need against Out-of-Distribution Overconfidence

Keke Tang[1*], **Chao Hou**[1*], Weilong Peng[1], Xiang Fang[2],
Zhize Wu[3], Yongwei Nie[4], Wenping Wang[5], Zhihong Tian[1]

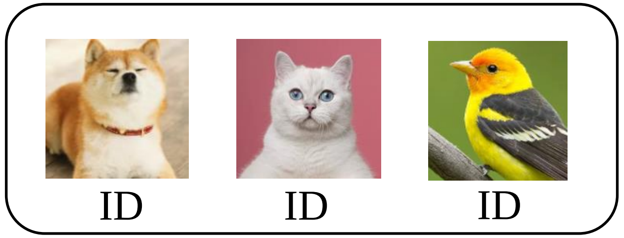
Deficiencies of DNN Closed World Assumption

Training Time

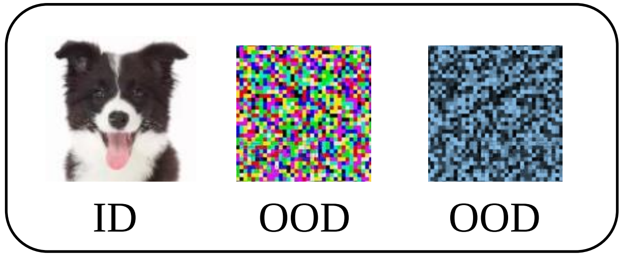


Dog/Cat/Bird Classifier

Test Time

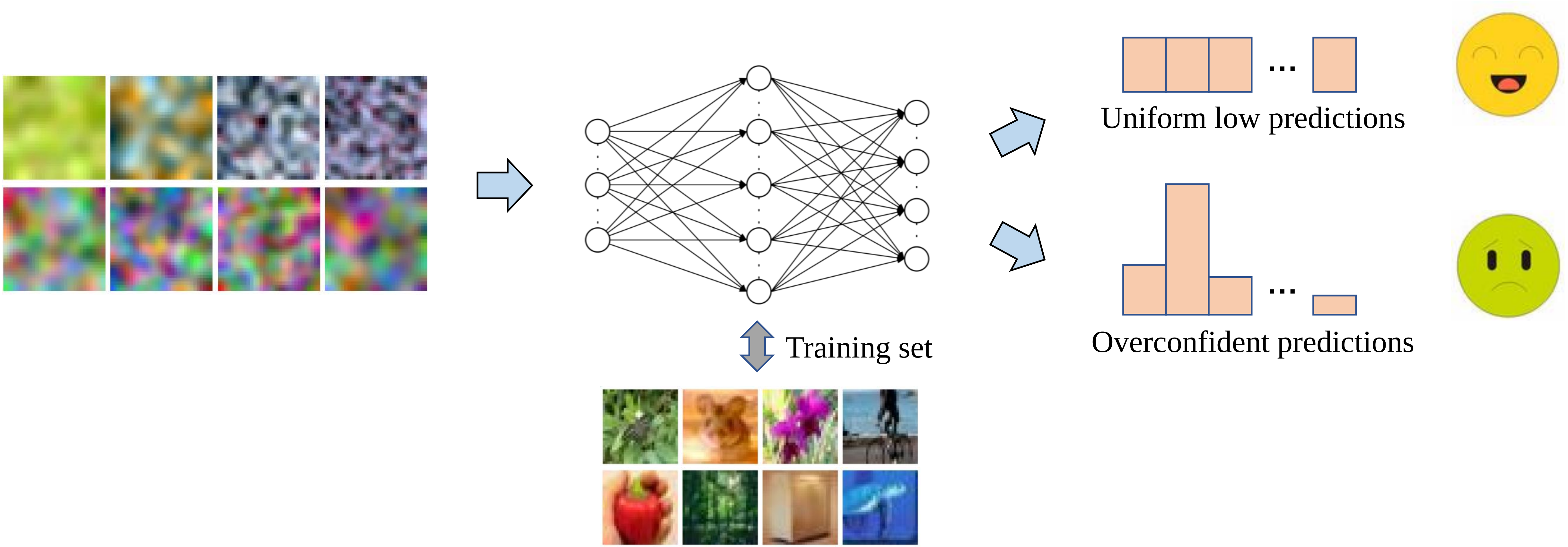


In the **closed world**, the test data has the same distribution as the training data.

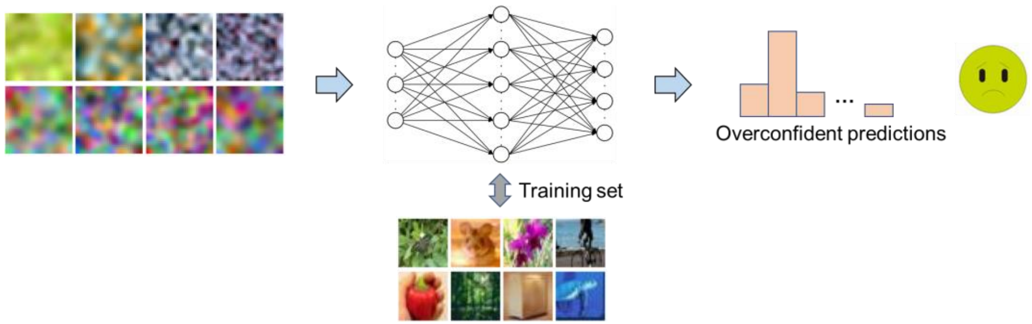


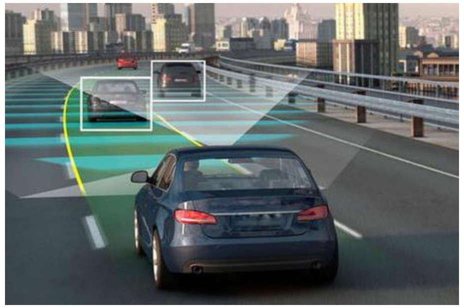
In the **open world**, there may exist test data with a distribution different from the training data.
(Unseen、Unknown)

OOD Overconfidence Issue



Influence of OOD Overconfidence Issue





Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem

Proved to be inevitable

Matthias Hein
University of Tübingen

Maksym Andriushchenko
Saarland University

Julian Bitterwolf
University of Tübingen

Related Work

OOD detection

Designing Detection Score

$$G_i(\mathbf{x};\theta_i) = \begin{cases} \text{in}, & \text{if } S_i(\mathbf{x};\theta_i) \geq \gamma_i \\ \text{out}, & \text{if } S_i(\mathbf{x};\theta_i) < \gamma_i \end{cases}$$

MSP [Hendrycks et al. 2016]

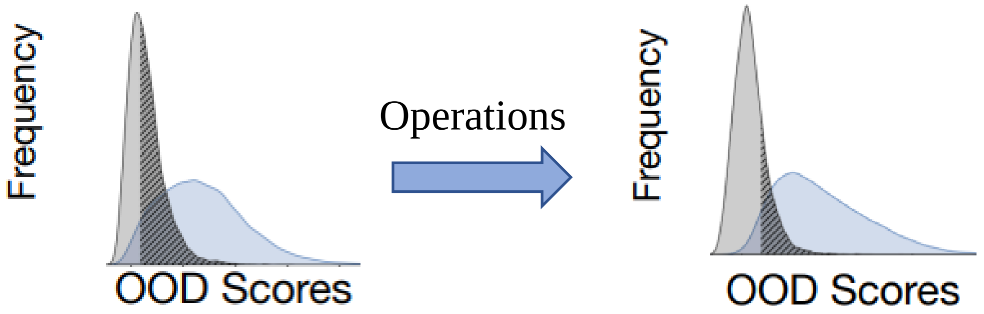
Mahalanobis metrics [Lee et al. 2018]

Energy [Liu et al. 2020]

FeatureNorm [Yu et al. 2023]

⋮

Widen the Differentiation



ODIN [Liang et al. 2018]

ReAct [Sun et al. 2021]

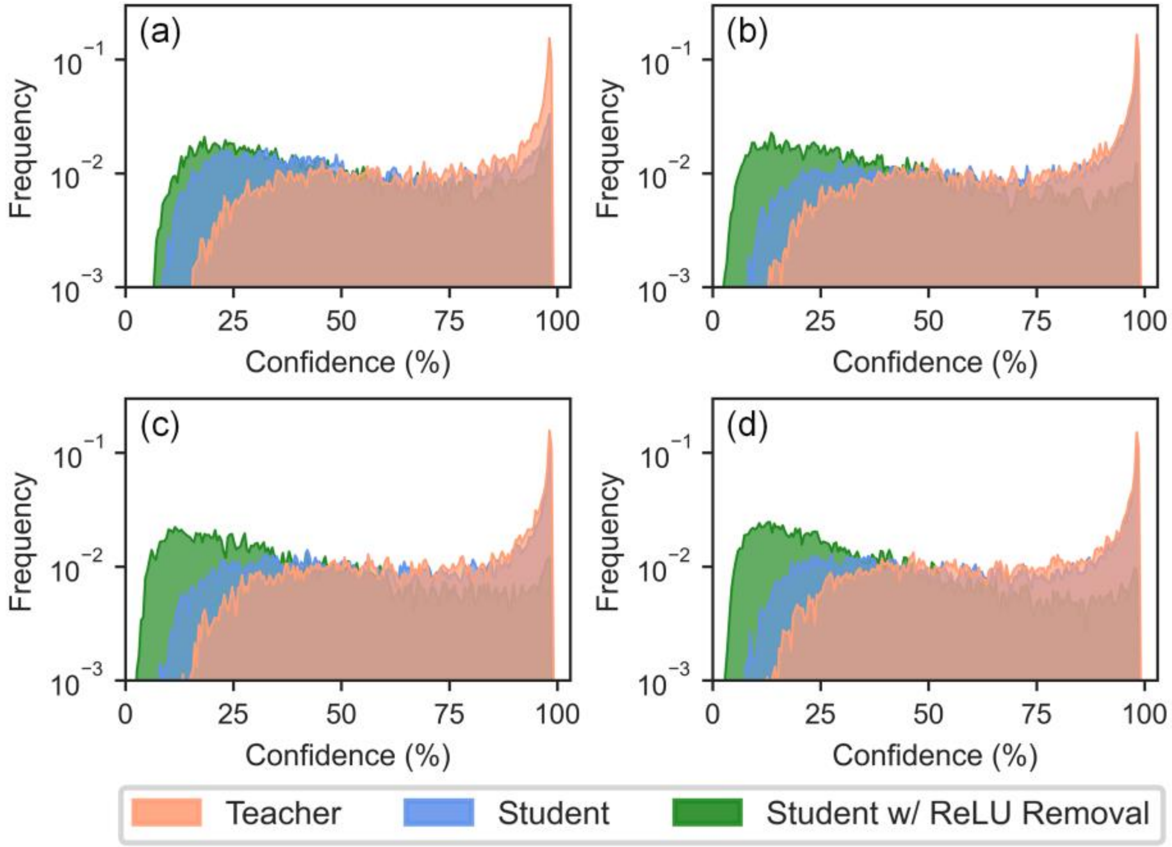
DICE [Sun et al. 2022]

BATS [Zhu et al. 2022]

⋮

Motivation

DNNs can indeed be regarded as **complex systems**, often characterized by **over-parameterization**. This excessive complexity can cause DNNs to exhibit emergent behaviors, leading to unpredictable outcomes, e.g., **OOD overconfidence Issue**.



How Complexity Affects Overconfidence?

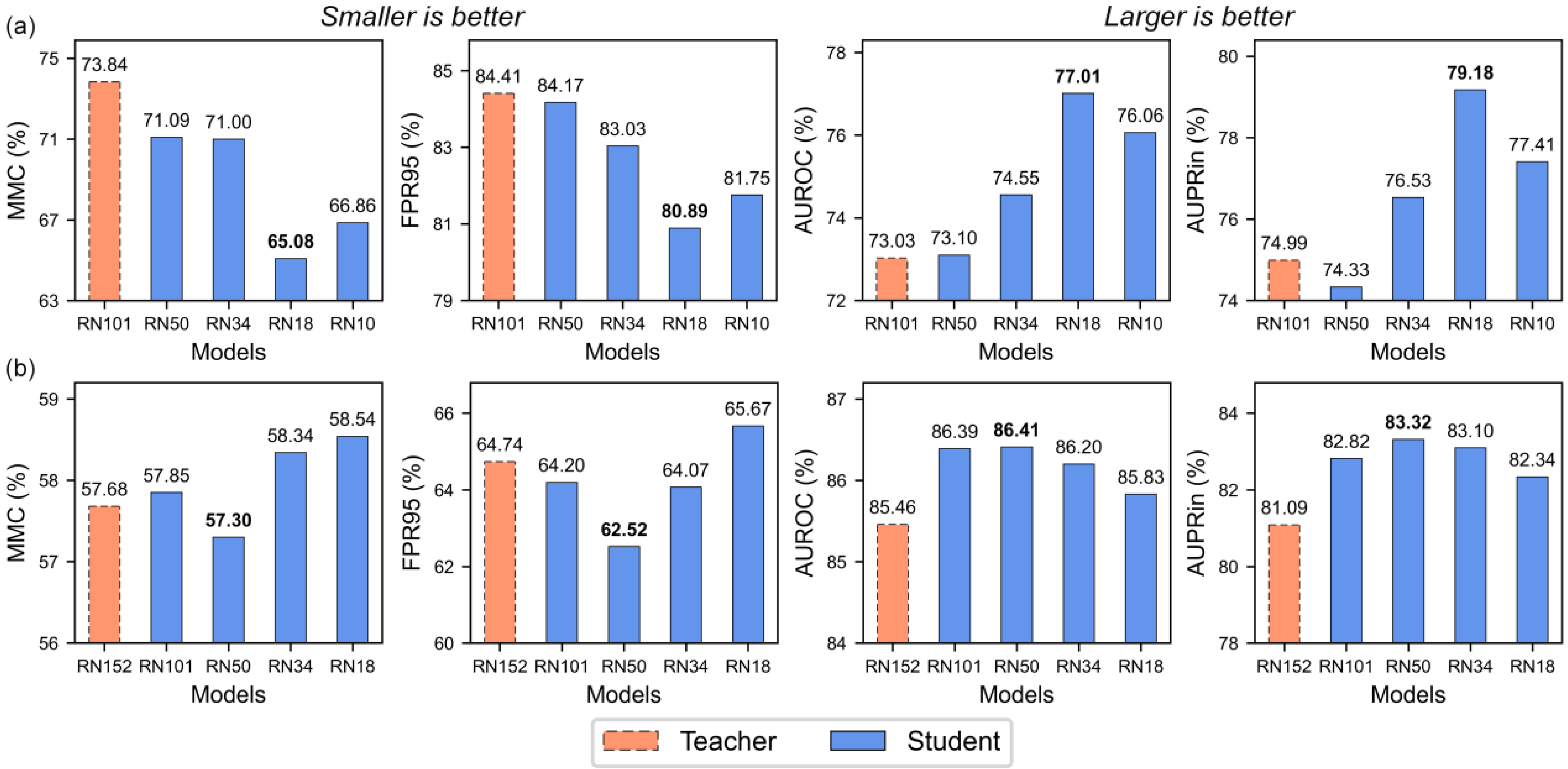
Capacity - Effectiveness of Knowledge Distillation

Table 1. Comparison of MMC and OOD detection performance for student models (ResNet-18) distilled from a teacher model (ResNet-50) using different knowledge distillation methods on CIFAR datasets. Results are averaged across multiple OOD datasets.

ID	CIFAR-10			CIFAR-100		
Method	MMC↓	FPR95↓	AUROC↑	MMC↓	FPR95↓	AUROC↑
Teacher	79.58	54.23	91.78	72.12	83.67	74.73
KD [22]	**76.81**	49.81	92.42	62.69	79.77	76.31
FitNet [37]	79.46	50.58	92.26	60.60	77.53	77.78
AT [64]	77.90	**48.09**	**92.69**	64.11	82.14	75.51
SP [48]	78.81	50.52	92.22	**58.75**	**77.12**	**80.01**

How Complexity Affects Overconfidence?

Capacity - Analysis of Student Networks with Varying Capacities



How Complexity Affects Overconfidence?

Nonlinearity - Analysis of Different Removal Proportions

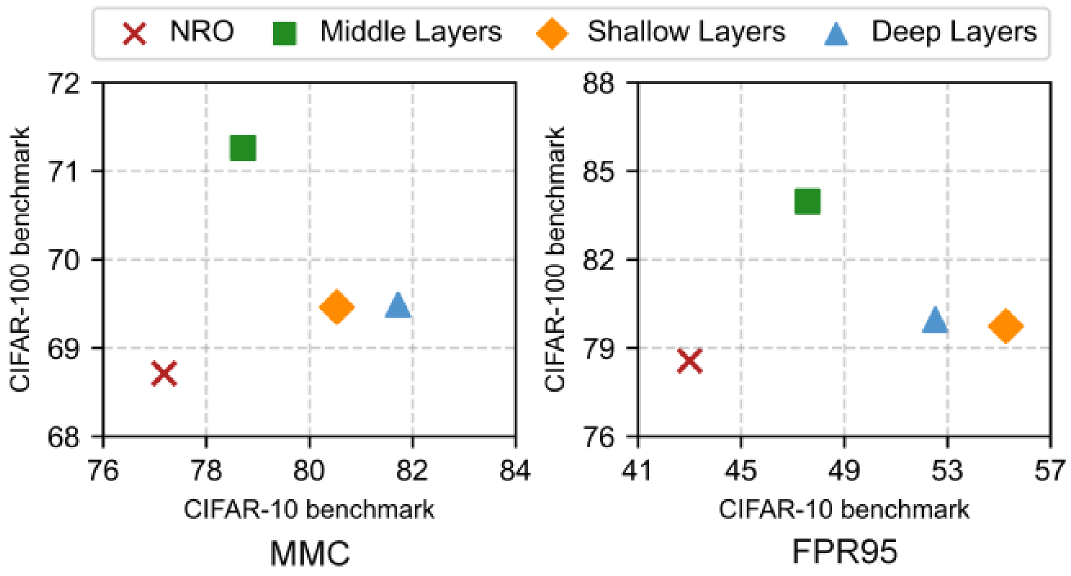
Table 2. Comparison of MMC and OOD detection performance for different ReLU removal proportions using ResNet-101 as the classifier. Results are averaged across multiple OOD datasets.

ID	CIFAR-10			CIFAR-100		
Ratio	MMC↓	FPR95↓	AUROC↑	MMC↓	FPR95↓	AUROC↑
0%	83.46	51.92	92.53	73.84	84.41	73.03
1%	84.46	57.58	90.48	69.14	82.43	76.22
2%	**77.18**	**43.00**	**93.69**	72.26	83.07	75.25
3%	77.91	46.93	93.06	69.37	78.74	78.12
5%	83.07	56.53	90.71	**68.71**	**78.57**	**79.43**
10%	79.60	50.06	92.25	71.24	80.94	77.70
20%	80.78	50.31	92.30	73.54	82.73	74.81
50%	79.22	48.34	92.73	70.40	82.22	75.40

How Complexity Affects Overconfidence?

Nonlinearity

Analysis of Different Removal Locations



$$\mathrm{NRO} = \frac{\text{Number of negative convolutional responses}}{\text{Total number of convolutional responses}}.$$

Impact of ReLU Removal Initiation Points

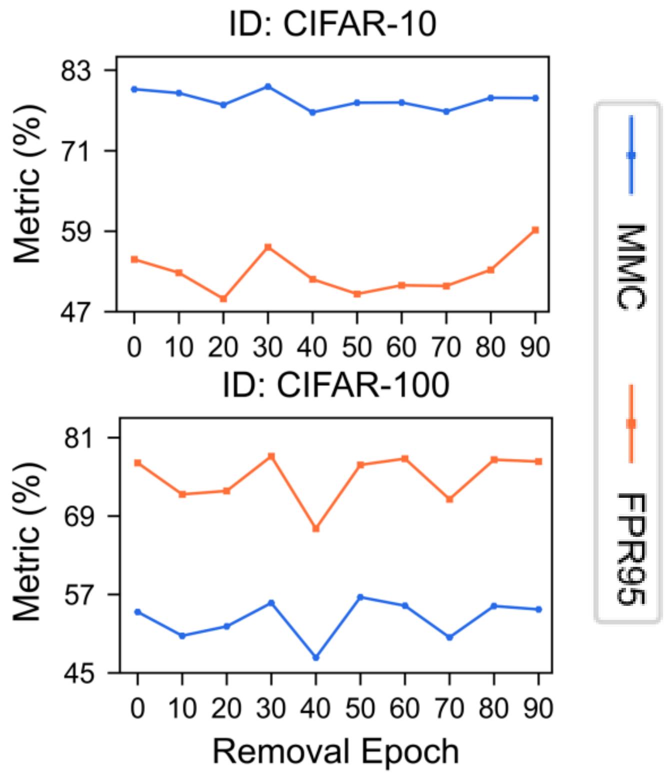
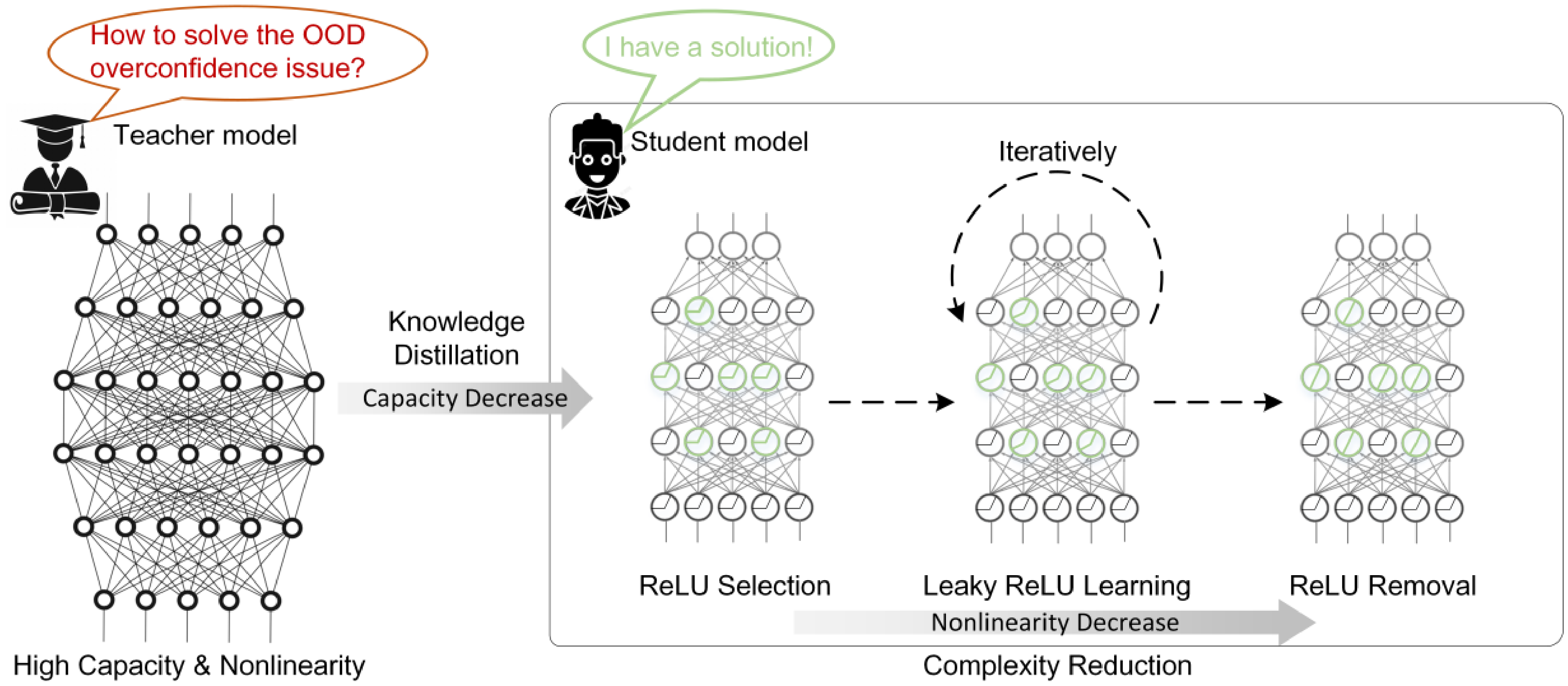


Illustration of Our Model Simplification Pipeline

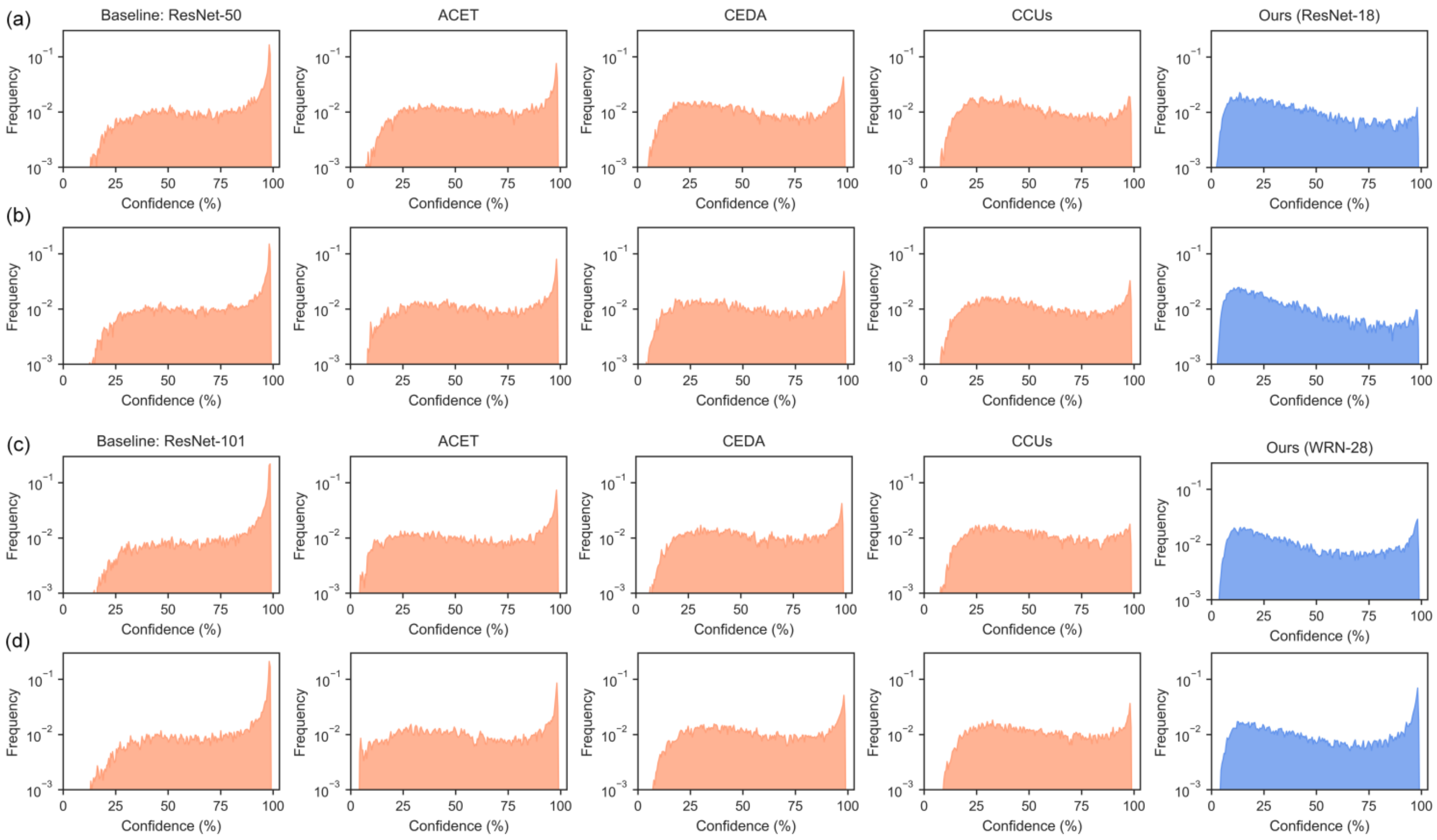


Given a complex teacher model with high capacity and nonlinearity, we use knowledge distillation to transfer information to a low-capacity student model. During distillation, ReLU operations are selectively replaced with Leaky ReLU, and the negative slope is iteratively adjusted from 0 to 1, ultimately transforming ReLU into an identity function.

Results of Mitigating OOD Overconfidence

Model	Method	ID: CIFAR-10						ID: CIFAR-100					
		SVHN	LSUN-R	iSUN	Textures	TINR	Average	SVHN	LSUN-R	iSUN	Textures	TINR	Average
ResNet-50	—	75.34	79.64	79.86	81.53	81.53	79.58	72.50	72.07	72.20	72.74	71.11	72.12
	CEDA	74.73	70.91	72.02	73.37	78.64	73.93	55.64	54.77	56.25	62.85	55.42	56.99
	ACET	74.29	67.35	67.72	70.37	69.87	69.92	63.68	62.49	62.19	64.44	61.15	62.79
	CCUs	69.56	63.51	63.10	68.64	68.92	66.75	57.33	52.30	54.71	61.25	54.89	56.09
	Ours (ResNet-18)	78.24	73.29	74.37	79.51	76.98	76.48	54.75	**43.65**	**42.27**	61.40	**39.69**	**48.35**
	Ours (WRN-16)	63.65	66.32	68.49	77.09	70.18	69.15	**43.06**	51.76	51.99	**57.96**	52.37	51.43
	CEDA + Ours (WRN-16)	**61.31**	67.40	69.64	70.09	73.54	68.40	47.84	59.40	60.69	59.24	60.00	57.43
	ACET + Ours (WRN-16)	67.44	61.28	62.36	70.05	**65.48**	65.32	52.54	59.45	61.59	60.82	60.48	58.98
	CCUs + Ours (WRN-16)	65.97	**58.81**	**61.32**	**64.84**	65.96	**63.38**	50.88	49.04	51.85	58.40	51.82	52.40
ResNet-101	—	83.44	82.13	83.02	84.56	84.17	83.46	65.25	77.21	78.03	73.38	75.32	73.84
	CEDA	70.07	68.26	68.42	74.42	74.02	71.04	58.82	58.18	59.00	66.23	58.47	60.14
	ACET	75.82	53.69	55.00	**64.50**	59.98	61.80	63.62	59.57	58.44	60.34	58.01	60.00
	CCUs	73.56	65.72	64.98	66.94	69.83	68.21	58.45	54.53	56.52	59.80	56.32	57.12
	Ours (ResNet-34)	75.22	74.81	75.94	82.08	78.86	77.38	63.62	58.42	60.13	66.69	59.89	61.75
	Ours (WRN-28)	71.21	74.73	76.03	81.80	77.89	76.33	55.69	**47.84**	**51.20**	63.69	**54.09**	54.50
	CEDA + Ours (WRN-28)	67.15	68.21	70.59	76.77	73.83	71.31	56.51	65.12	66.59	62.51	66.77	63.50
	ACET + Ours (WRN-28)	**59.33**	**52.70**	**53.07**	66.95	**58.05**	**58.02**	56.10	62.48	63.94	58.34	63.86	60.95
	CCUs + Ours (WRN-28)	67.79	61.74	64.20	65.95	68.80	65.70	**53.71**	50.96	53.58	**57.86**	54.52	**54.13**

Results of Mitigating OOD Overconfidence



Histograms (logarithmic scale) of MMC values for (a, b) ResNet-50 and (c, d) ResNet-101, w/ and w/o applying mitigation techniques: ACET, CEDA, CCUs, and Ours. All models are trained on CIFAR-100 and evaluated on OOD datasets: (a, c) LSUN-R and (b, d) TinyImageNet-R. In our approach, ResNet-18 and WRN-28 serve as students for ResNet-50 and ResNet-101, respectively.

Results of Enhancing OOD Detection

ID	Model	Method	OOD										Average	
			SVHN		LSUN-R		iSUN		Textures		TINR			
			FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIFAR-10	ResNet-50	MSP	43.75	94.19	54.32	91.96	55.46	91.66	59.01	90.31	58.61	90.80	54.23	91.78
		MSP + Ours (ResNet-18)	51.82	93.27	41.93	93.89	44.86	93.27	**55.64**	**90.57**	50.18	91.97	48.89	92.59
		MSP + Ours (WRN-16)	**28.18**	**96.12**	**35.74**	**95.01**	**39.78**	**94.34**	57.00	89.79	**42.35**	**93.75**	**40.61**	**93.80**
		MaxLogit	22.07	96.31	32.03	94.95	34.95	94.53	46.33	91.39	44.96	92.57	36.07	93.95
		MaxLogit + Ours (ResNet-18)	30.25	94.94	25.06	95.34	27.91	94.73	**44.50**	**91.71**	36.75	92.73	32.89	93.89
		MaxLogit + Ours (WRN-16)	**17.61**	**96.96**	**18.54**	**96.92**	**21.69**	**96.39**	47.70	89.58	**27.81**	**95.55**	**26.67**	**95.08**
		Energy	20.35	96.40	29.79	95.11	32.80	94.67	45.41	**91.44**	43.76	92.64	34.42	94.05
		Energy + Ours (ResNet-18)	28.53	95.07	23.99	95.43	26.55	94.83	**44.13**	89.70	35.92	92.79	31.82	93.56
		Energy + Ours (WRN-16)	**18.80**	**96.88**	**18.07**	**97.02**	**20.91**	**96.50**	47.89	89.54	**27.63**	**95.61**	**26.66**	**95.11**
		ReAct	37.66	91.10	**16.24**	96.64	**19.65**	95.84	46.72	86.29	30.79	93.46	30.21	92.67
		ReAct + Ours (ResNet-18)	31.59	94.62	23.87	95.51	26.78	94.86	**43.60**	**90.19**	35.87	92.92	32.34	93.62
		ReAct + Ours (WRN-16)	**19.07**	**96.84**	17.90	**97.03**	20.66	**96.51**	47.45	89.84	**27.24**	**95.68**	**26.46**	**95.18**
		FeatureNorm	9.25	97.97	77.34	82.07	70.29	83.96	48.97	84.16	67.71	84.49	54.71	86.53
		FeatureNorm + Ours (ResNet-18)	**2.73**	**99.35**	24.34	95.75	**21.04**	96.28	**22.82**	**94.89**	29.52	94.36	**20.09**	**96.13**
		FeatureNorm + Ours (WRN-16)	3.37	99.26	**22.03**	**96.43**	21.54	**96.47**	53.67	78.85	**29.35**	**94.50**	25.99	93.10
CIFAR-100	ResNet-101	MSP	76.30	81.14	86.67	69.76	88.11	69.06	85.69	73.73	85.27	71.47	84.41	73.03
		MSP + Ours (ResNet-34)	77.48	78.92	72.74	83.09	74.83	82.14	**82.50**	**77.26**	74.52	**82.20**	76.41	80.72
		MSP + Ours (WRN-28)	**74.87**	**81.39**	**62.57**	**85.71**	**66.94**	**83.93**	83.28	76.33	**69.44**	81.70	**71.42**	**81.81**
		MaxLogit	**68.05**	**86.77**	82.84	74.34	84.94	73.40	87.25	74.81	81.92	75.05	81.00	76.87
		MaxLogit + Ours (ResNet-34)	71.69	83.47	63.74	87.26	65.69	86.54	**81.74**	**78.46**	66.88	**86.44**	69.95	**84.43**
		MaxLogit + Ours (WRN-28)	73.84	82.61	**54.30**	**89.11**	**60.29**	**87.26**	83.62	77.06	**64.12**	84.98	**67.23**	84.20
		Energy	**68.38**	**86.94**	82.38	74.64	84.21	73.67	88.03	74.62	81.40	75.25	80.88	77.02
		Energy + Ours (ResNet-34)	69.20	84.14	58.83	88.05	61.10	87.37	**81.68**	**78.51**	62.89	**87.18**	66.74	**85.05**
		Energy + Ours (WRN-28)	74.17	82.61	**48.60**	**90.07**	**55.05**	**88.12**	84.10	77.01	**60.41**	85.71	**64.47**	84.70
		ReAct	65.68	88.14	86.44	69.18	86.97	69.09	83.44	80.13	83.80	71.59	81.27	75.63
		ReAct + Ours (ResNet-34)	**60.26**	**88.30**	57.94	86.49	58.11	86.52	**63.07**	**85.78**	60.03	**85.69**	**59.88**	**86.56**
		ReAct + Ours (WRN-28)	71.89	86.73	**47.34**	**89.20**	**53.66**	**87.46**	80.57	81.48	**58.97**	84.66	62.49	85.91
		FeatureNorm	**31.88**	**94.59**	99.54	34.11	99.10	39.02	**53.42**	**82.32**	98.22	42.32	76.43	58.47
		FeatureNorm + Ours (ResNet-34)	41.47	90.47	92.46	71.49	91.41	72.89	65.18	80.33	89.99	69.00	76.10	76.84
		FeatureNorm + Ours (WRN-28)	54.17	89.08	**80.86**	**79.52**	**80.69**	**79.91**	71.49	74.37	**78.00**	**80.74**	**73.04**	**80.72**

Results of Enhancing OOD Detection

Model	Method	OOD								Average	
		SUN		Places		Textures		OpenImage-O			
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
DenseNet-201	MSP	**65.76**	**82.24**	69.53	81.14	67.82	79.36	64.92	84.47	67.01	81.80
	MSP + Ours (DenseNet-161)	66.31	81.72	**68.61**	**81.32**	**66.97**	**79.55**	**62.12**	**85.10**	**66.00**	**81.92**
	MaxLogit	55.44	86.06	61.89	83.99	58.48	83.45	**59.64**	86.94	58.86	85.11
	MaxLogit + Ours (DenseNet-161)	**55.33**	**87.67**	**59.95**	**86.26**	**55.80**	**85.84**	60.03	**88.60**	**57.78**	**87.09**
	Energy	52.87	86.04	60.46	83.71	56.13	83.35	59.87	86.65	57.33	84.94
	Energy + Ours (DenseNet-161)	**50.88**	**88.25**	**56.90**	**86.60**	**52.25**	**86.45**	**58.76**	**88.60**	**54.70**	**87.48**
	ReAct	**46.00**	**90.81**	57.46	86.11	55.14	84.19	69.32	74.43	56.98	83.89
	ReAct + Ours (DenseNet-161)	49.21	88.55	**55.02**	**86.94**	**50.99**	**86.75**	**57.61**	**88.64**	**53.21**	**87.72**
	FeatureNorm	43.06	90.05	55.71	85.42	**50.46**	**79.90**	**79.97**	**67.22**	57.30	80.65
	FeatureNorm + Ours (DenseNet-161)	**36.52**	**91.79**	**49.24**	**87.50**	56.73	78.29	82.20	65.49	**56.17**	**80.77**
WRN-101	MSP	**66.93**	82.44	70.48	81.35	**65.14**	80.91	**59.87**	86.13	**65.61**	82.71
	MSP + Ours (WRN-50)	67.75	**82.80**	**69.07**	**81.81**	66.38	**80.96**	60.91	**86.78**	66.03	**83.09**
	MaxLogit	63.25	84.75	68.40	82.69	57.39	84.38	55.84	88.43	61.22	85.06
	MaxLogit + Ours (WRN-50)	**61.92**	**86.18**	**64.34**	**84.40**	**56.74**	**85.60**	**55.69**	**89.25**	**59.67**	**86.36**
	Energy	65.06	84.46	70.32	82.27	57.11	84.37	58.04	88.02	62.63	84.78
	Energy + Ours (WRN-50)	**63.18**	**85.96**	**66.06**	**84.00**	**55.90**	**85.68**	**57.87**	**88.77**	**60.75**	**86.10**
	ReAct	52.22	88.11	58.76	85.47	57.09	84.99	**49.63**	**87.82**	54.43	86.60
	ReAct + Ours (WRN-50)	**37.02**	**92.24**	**48.54**	**88.70**	**50.90**	**87.36**	54.49	85.92	**47.74**	**88.56**
	FeatureNorm	70.07	78.89	78.61	73.23	20.98	94.79	75.09	74.81	61.19	80.43
	FeatureNorm + Ours (WRN-50)	**60.22**	**83.67**	**74.01**	**75.96**	**17.71**	**96.09**	**69.94**	**78.90**	**55.47**	**83.66**

Importance of Knowledge Distillation and ReLU Removal

Table 6. Comparison of MMC and OOD detection performance for ResNet-50 enhanced by Ours (ResNet-18) and its variants w/o knowledge distillation (KD) and ReLU removal (RR), using CIFAR-10 as the ID dataset, and SVHN, LSUN-R, iSUN, Textures, and TINR as the OOD datasets. Results are averaged.

KD	RR	MMC↓	FPR95↓	AUROC↑
		79.58	54.23	91.78
✓		76.81	49.81	92.42
	✓	78.46	50.50	91.76
✓	✓	**76.48**	**48.89**	**92.59**

Knowledge Distillation vs. Network Pruning for Capacity Reduction

Table 7. Comparison of MMC and OOD detection performance for ResNet-50 and ResNet-101 reduced to comparable parameter counts via KD [22] and Prune [14]. CIFAR-10 serves as the ID dataset, with SVHN, LSUN-R, iSUN, Textures, and TINR as OOD datasets. Results are averaged across all five OOD datasets

Model	Capacity Reduction	MMC↓	FPR95↓	AUROC↑
ResNet-50	—	79.58	54.23	91.78
	Prune	77.34	52.55	92.27
	KD (ResNet-18)	**76.81**	**49.81**	**92.42**
ResNet-101	—	83.46	51.92	92.53
	Prune	78.20	50.92	91.61
	KD (WRN-28)	**77.01**	**41.09**	**92.56**

Thanks!