Diversity Matters When Learning From Ensembles

Giung Nam[1*] Jongmin Yoon[1*] Yoonho Lee[2,3] Juho Lee[1,2]

(*Equal contribution)

[1]KAIST, South Korea

[2]AITRICS, South Korea

[3]Stanford University, USA

NeurIPS 2021

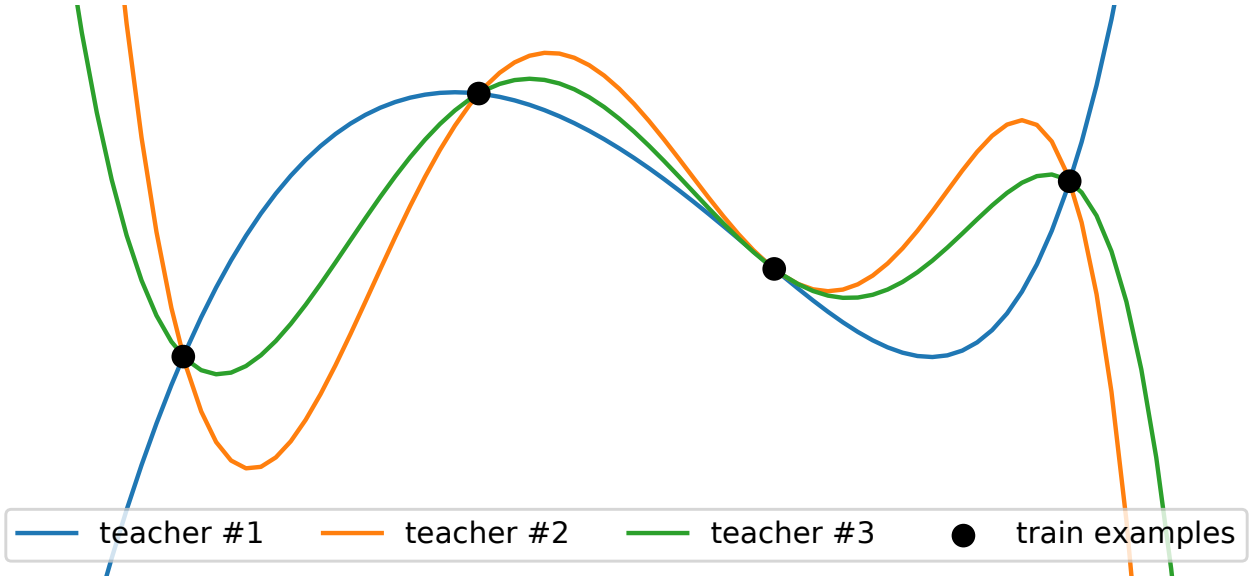
Learning From Ensembles

- Deep Ensemble (DE):

$$p(y|\boldsymbol{x}) = \frac{1}{M}\sum_{m=1}^{M} p_{\boldsymbol{\theta}_m}(y|\boldsymbol{x}, \boldsymbol{\theta}_m).$$

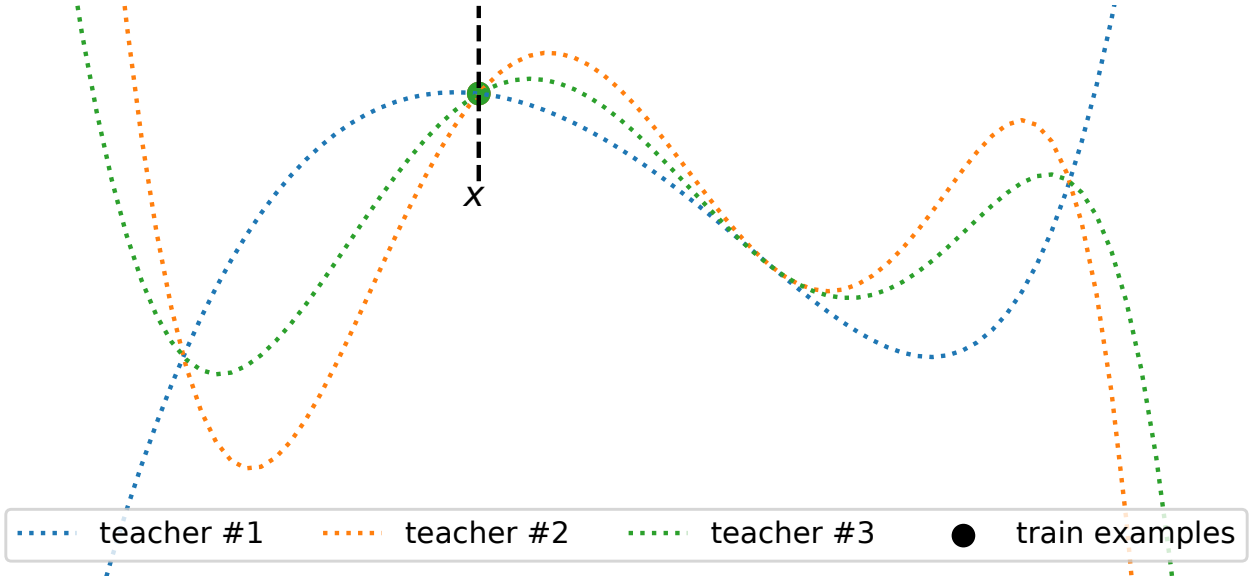
- How to learn from DE? $\rightarrow$ Knowledge Distillation (KD).

Motivating Example



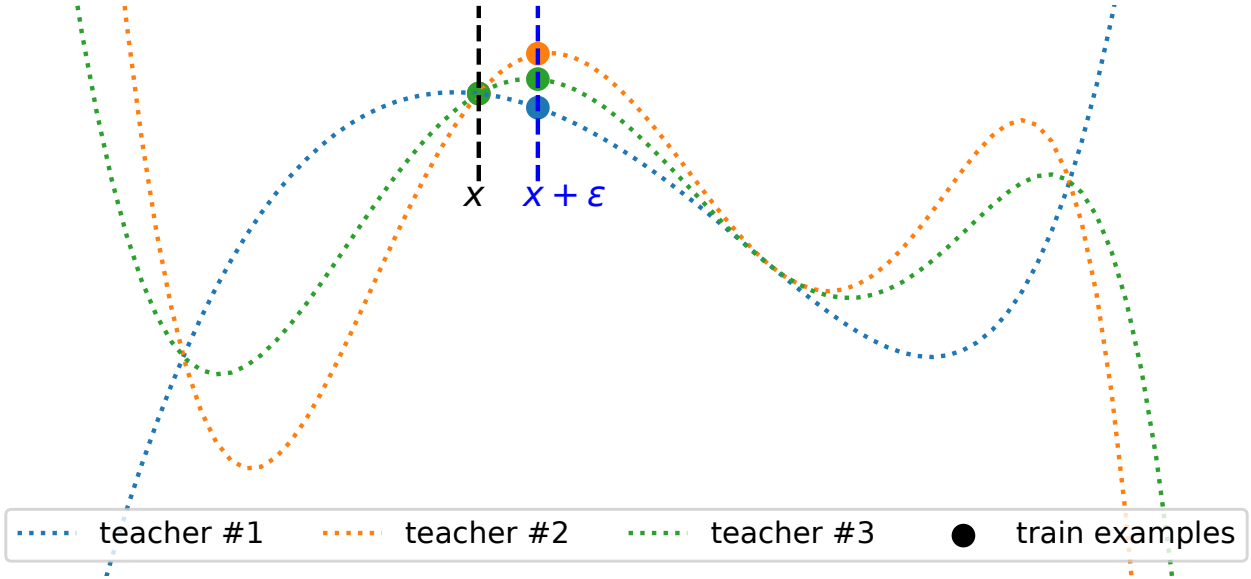
- Assume that we have three pre-trained teacher models.

Motivating Example



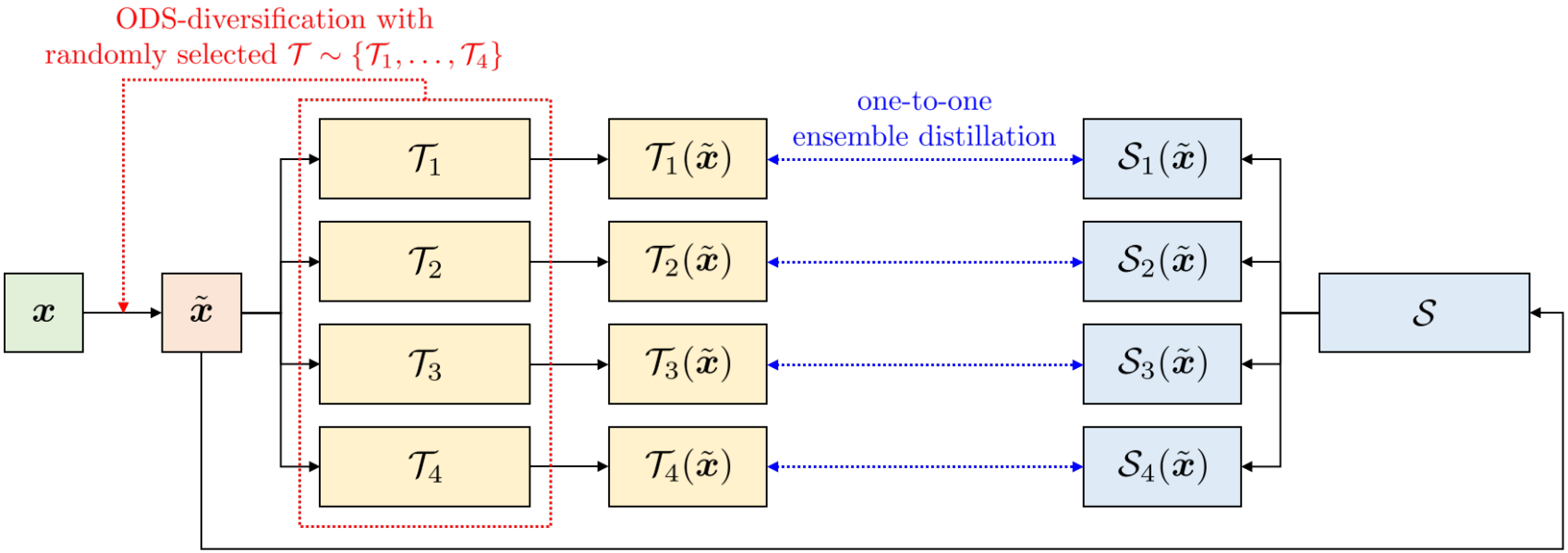
- Assume that we have three pre-trained teacher models.
- We cannot capture *diversity* if we reuse the train data x.

Motivating Example



- Assume that we have three pre-trained teacher models.
- We cannot capture *diversity* if we reuse the train data x.
- Idea: use perturbed data $x + \varepsilon$ that can capture *diversity*!

Our Approach



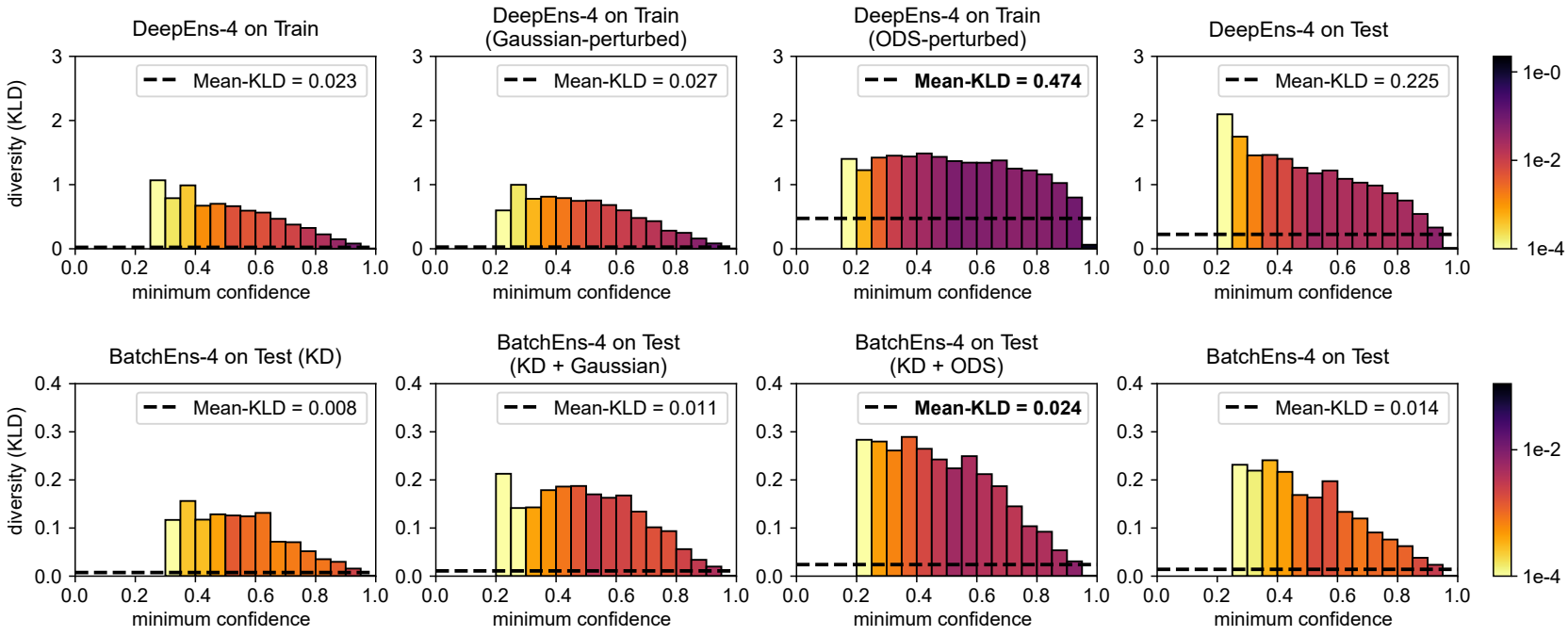
- Output diversified sampling (ODS):

$$\boldsymbol{\varepsilon}_{\mathsf{ODS}}(\boldsymbol{x}, \mathcal{F}, \boldsymbol{w}) = \frac{\nabla_{\boldsymbol{x}}(\boldsymbol{w}^\top \mathcal{F}(\boldsymbol{x}))}{\|\nabla_{\boldsymbol{x}}(\boldsymbol{w}^\top \mathcal{F}(\boldsymbol{x}))\|_2} \in \mathbb{R}^D,$$

where $\boldsymbol{w} \sim \mathrm{Unif}([-1, 1])^K$.

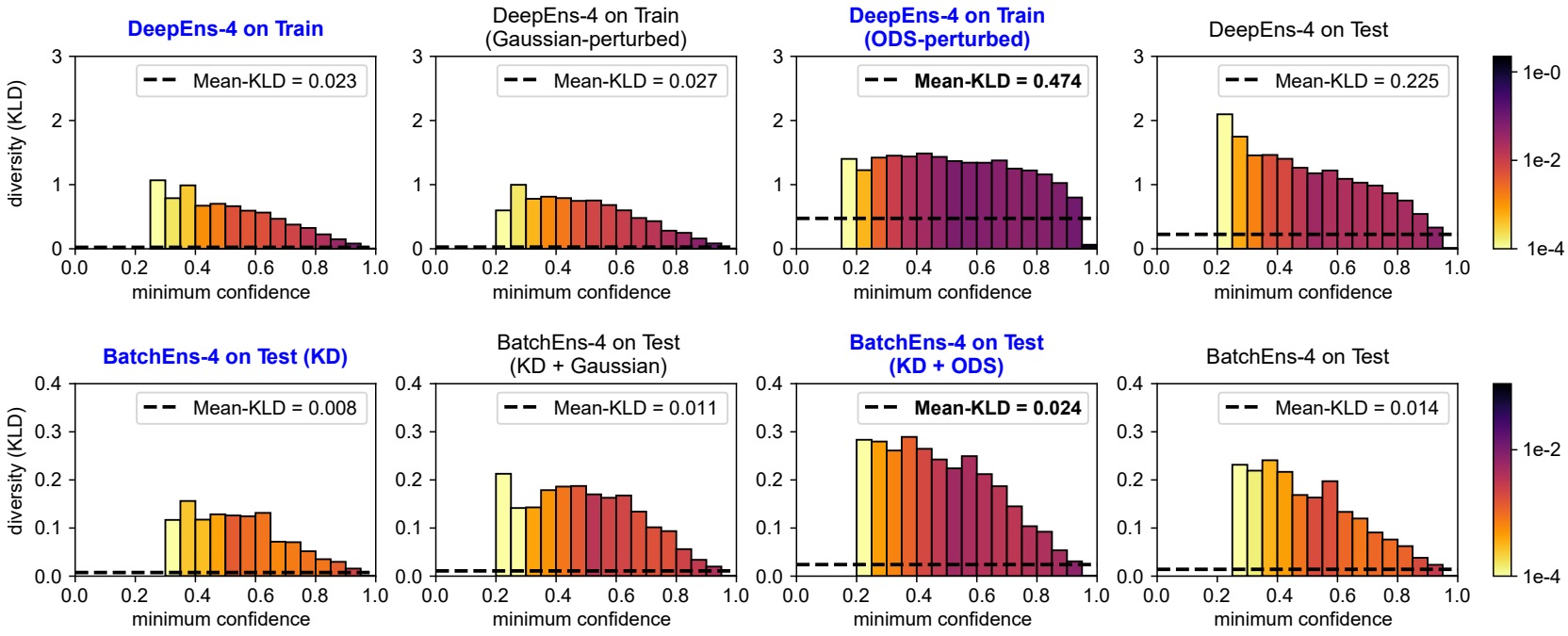
Experimental Results

Impact of ODS on diversities



Experimental Results

Impact of ODS on diversities



Evaluation Results

Image classification tasks

Method	# Params	ACC	NLL	BS	ECE	DEE
		ResNet-32 on CIFAR-10				
$\mathcal{T}$: DeepEns-4	1.86 M	94.42	0.167	0.082	0.008	-
$\mathcal{S}$: BatchEns-4	0.47 M	93.37±0.11	0.204±0.002	0.099±0.001	0.008±0.001	1.419±0.075
+ KD		93.98±0.20	0.188±0.003	0.091±0.002	0.009±0.002	2.019±0.174
+ KD + Gaussian		93.93±0.12	0.187±0.001	0.090±0.001	0.009±0.002	2.042±0.089
+ KD + ODS		93.89±0.10	0.181±0.002	0.090±0.001	**0.006**±0.001	2.486±0.164
+ KD + ConfODS		**94.01**±0.19	**0.180**±0.001	**0.089**±0.001	0.007±0.001	**2.524**±0.080
$\mathcal{T}$: DeepEns-8	3.71 M	94.78	0.157	0.077	0.005	-
$\mathcal{S}$: BatchEns-8	0.48 M	93.47±0.14	0.202±0.005	0.098±0.002	**0.006**±0.002	1.494±0.156
+ KD		94.15±0.13	0.182±0.001	0.088±0.001	0.010±0.001	2.391±0.113
+ KD + Gaussian		94.09±0.08	0.184±0.002	0.089±0.001	0.010±0.002	2.206±0.175
+ KD + ODS		94.13±0.08	0.175±0.003	**0.086**±0.001	**0.006**±0.002	2.991±0.322
+ KD + ConfODS		**94.18**±0.12	**0.174**±0.002	**0.086**±0.001	0.007±0.001	**3.064**±0.246
		WRN28x10 on CIFAR-100				
Method	# Params	ACC	NLL	BS	ECE	DEE
$\mathcal{T}$: DeepEns-4	146.15 M	82.52	0.661	0.247	0.022	-
$\mathcal{S}$: BatchEns-4	36.62 M	80.34±0.08	0.755±0.007	0.280±0.002	0.027±0.001	1.449±0.085
+ KD		80.51±0.22	0.744±0.003	0.274±0.001	**0.021**±0.003	1.582±0.035
+ KD + Gaussian		80.39±0.12	0.761±0.006	0.277±0.000	0.022±0.003	1.379±0.072
+ KD + ODS		**81.88**±0.32	0.674±0.016	0.257±0.006	0.026±0.003	3.303±0.769
+ KD + ConfODS		81.85±0.32	**0.672**±0.010	**0.256**±0.003	0.024±0.000	**3.333**±0.491

Evaluation Results

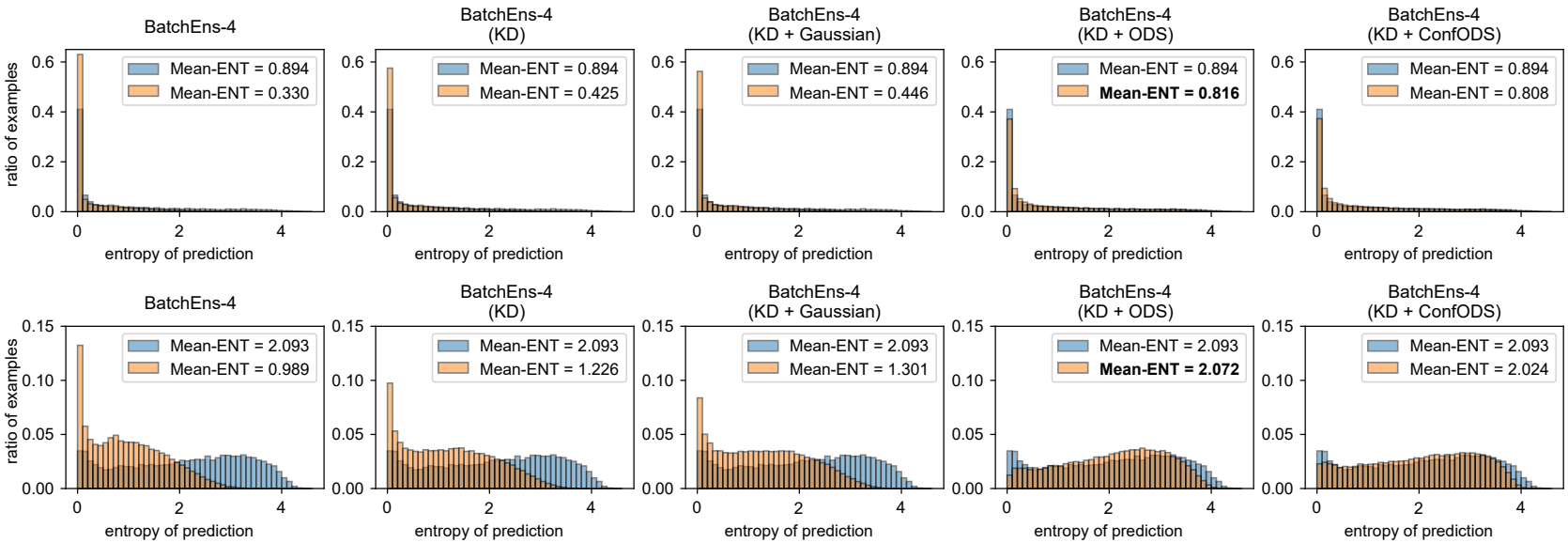
Cross-architecture knowledge distillation for compression

		WRN28x5 on CIFAR-100			
Method	# Params	ACC	NLL	BS	ECE
$\mathcal{T}$: DeepEns-4 (WRN28x10)	146.15 M	82.52	0.661	0.247	0.022
$\mathcal{S}$: BatchEns-4	9.20 M	78.75$\pm$0.11	0.801$\pm$0.012	0.297$\pm$0.003	0.021$\pm$0.002
+ KD		78.89$\pm$0.10	0.804$\pm$0.012	0.296$\pm$0.003	0.022$\pm$0.001
+ KD + Gaussian		78.80$\pm$0.41	0.815$\pm$0.009	0.297$\pm$0.005	**0.020**$\pm$0.002
+ KD + ODS		80.24$\pm$0.05	0.742$\pm$0.008	0.279$\pm$0.002	0.028$\pm$0.004
+ KD + ConfODS		**80.62**$\pm$0.25	**0.733**$\pm$0.007	**0.275**$\pm$0.003	0.027$\pm$0.001

		WRN28x5 on TinyImageNet			
Method	# Params	ACC	NLL	BS	ECE
$\mathcal{T}$: DeepEns-4 (WRN28x10)	146.40 M	69.90	1.242	0.403	0.016
$\mathcal{S}$: BatchEns-4	9.23 M	64.86	1.455	0.464	**0.022**
+ KD		65.86	1.432	**0.456**	0.022
+ KD + Gaussian		65.72	1.446	0.457	0.022
+ KD + ODS		**65.98**	**1.408**	**0.456**	0.022

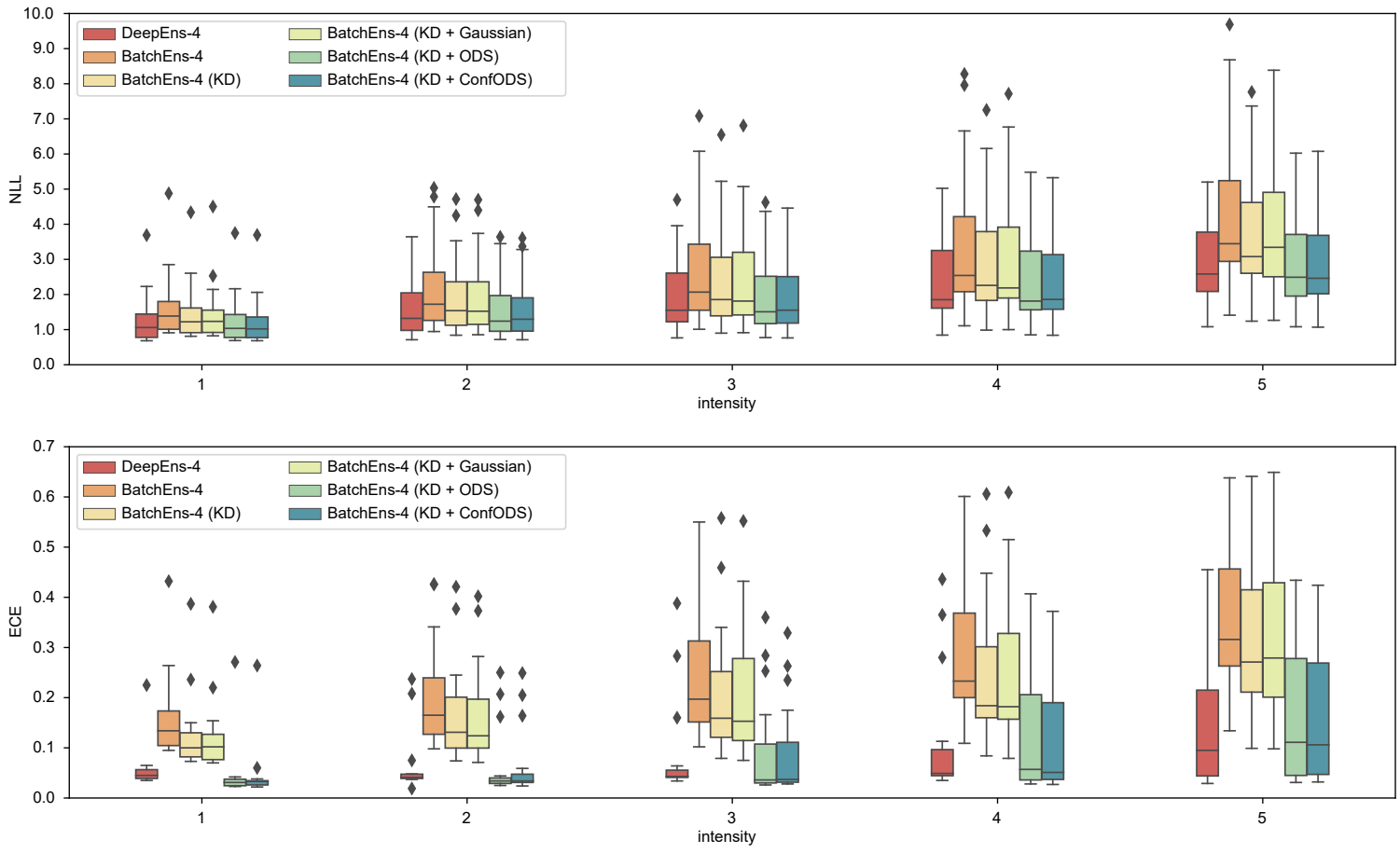
Evaluation Results

Predictive uncertainty for out-of-distribution examples



Evaluation Results

Calibration on corrupted dataset



Summary

- Diversity matters when learning from ensembles!
- ODS is one way to capture the *diversity* in KD framework.
- Such enhanced *diversities* help to reduce the performance gap between student and teacher in KD framework.