Table 1: Comprehensive comparison involving 15 different types of corruption from commonly-used domain adaption benchmark [52]. Substantial (≥ 0.5) improvement and degradation compared to the baseline MSP [6] are highlighted in blue or red respectively. DUL is the only method that achieves SOTA OOD detection performance without sacrificing generalization i.e., the value of the entire row is almost black or blue. The **best** or second best results are highlighted in bold or underlined. MD is the shorthand of Mahalanobis.

Method	$\mathcal{P}_{\text{train}}^{\text{ID}}$	$\mathcal{P}_{\text{train}}^{\text{SEM}}$	OOD generalization (Error rate ↓)															OOD detection	
			Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elast.	Pixel	JPEG	Avg.	FPR↓	AUC↑
MSP	CIFAR-10	None	77.0	67.7	74.5	36.1	78.5	40.7	41.4	40.5	46.7	31.4	22.6	40.3	35.3	42.8	44.9	48.0	42.0	89.3
EBM			77.0	67.7	74.5	36.1	78.5	40.7	41.4	40.5	46.7	31.4	22.6	40.3	35.3	42.8	44.9	48.0	32.5	89.3
Maxlogits			77.0	67.7	74.5	36.1	78.5	40.7	41.4	40.5	46.7	31.4	22.6	40.3	35.3	42.8	44.9	48.0	32.9	89.3
MD			77.0	67.7	74.5	36.1	78.5	40.7	41.4	40.5	46.7	31.4	22.6	40.3	35.3	42.8	44.9	48.0	32.5	93.9
Entropy		ImageNet-RC	81.7	72.9	87.0	38.2	88.9	42.2	44.6	45.5	53.9	33.1	24.4	39.8	37.2	47.6	48.8	52.4	6.6	98.7
EBM (FT)			82.3	73.5	90.6	39.2	90.7	41.4	44.4	47.3	54.9	34.3	24.7	**39.2**	37.7	51.5	49.3	53.4	3.6	98.4
DPN			77.6	68.6	83.6	39.7	86.8	43.3	44.9	47.0	53.7	36.4	26.1	43.5	37.6	47.5	**44.5**	52.0	4.3	98.5
POEM			83.7	76.6	88.0	44.1	90.1	44.3	46.4	52.8	60.8	39.5	28.9	43.4	42.5	56.1	54.5	56.8	**3.3**	**99.0**
WOODS			77.5	68.5	77.5	36.4	80.6	40.9	41.9	41.3	47.1	31.4	23.3	39.7	36.2	42.9	46.2	48.8	7.1	98.5
SCONE			**76.4**	**67.3**	75.3	36.2	**77.0**	40.9	41.6	40.2	**45.5**	31.4	23.1	40.1	35.9	**42.4**	45.4	47.9	7.0	98.5
DUL (Ours)			77.2	68.0	**75.1**	**35.8**	78.5	**39.7**	**40.6**	**39.8**	46.1	**31.0**	**22.4**	**39.2**	**34.8**	43.0	44.6	**47.7**	5.9	98.5
Entropy		TIN-597	83.3	75.4	79.8	38.0	82.7	42.2	44.2	44.5	53.4	32.8	23.7	**38.0**	37.9	44.6	62.9	52.2	11.6	97.9
EBM (FT)			81.2	72.8	78.1	37.6	80.0	42.9	43.6	43.7	51.1	33.5	23.9	40.2	38.1	45.1	73.3	52.3	19.4	87.5
DPN			81.8	72.9	79.1	39.6	81.3	44.9	46.2	45.4	52.4	34.3	25.6	40.9	39.2	47.7	76.4	53.8	17.3	94.9
POEM			78.9	70.3	74.7	38.6	**78.0**	43.4	44.3	43.0	50.2	34.6	25.4	43.1	37.3	45.4	81.6	52.6	34.3	86.8
WOODS			81.1	72.4	76.5	38.6	79.0	43.2	44.6	42.4	49.2	33.0	24.3	40.2	39.0	**41.2**	47.8	50.2	7.6	**98.3**
SCONE			80.9	72.3	77.2	37.9	78.7	42.4	43.4	42.8	49.5	32.2	24.0	39.3	38.0	41.8	48.3	50.0	8.0	98.2
DUL (Ours)			**77.0**	**67.7**	**74.2**	**36.2**	78.4	**40.8**	**41.3**	**40.2**	**46.3**	**31.4**	**22.9**	40.8	**35.3**	42.7	**45.1**	**48.0**	**6.9**	98.2
MSP	ImageNet-200	None	52.2	70.7	71.5	56.0	55.5	52.6	54.1	67.3	67.0	63.5	56.5	53.3	46.3	50.3	51.9	57.9	58.2	82.3
EBM			52.2	70.7	71.5	56.0	55.5	52.6	54.1	67.3	67.0	63.5	56.5	53.3	46.3	50.3	51.9	57.9	32.5	89.3
Maxlogits			52.2	70.7	71.5	56.0	55.5	52.6	54.1	67.3	67.0	63.5	56.5	53.3	46.3	50.3	51.9	57.9	51.9	88.2
Entropy		ImageNet-800	51.3	71.2	71.7	**54.4**	54.6	51.8	53.2	66.9	66.2	62.7	55.8	**51.4**	**45.1**	49.7	51.1	57.1	53.6	89.1
EBM (FT)			52.9	72.2	72.8	56.0	56.2	53.5	54.1	67.8	67.4	64.0	57.0	52.2	46.5	51.4	52.9	58.5	59.7	87.5
DPN			51.9	69.2	**69.6**	56.5	55.6	52.4	53.1	**65.5**	**65.0**	**62.2**	**55.5**	54.5	46.0	49.7	51.4	57.2	63.8	87.2
WOODS			51.4	69.4	70.0	55.2	54.8	51.9	52.9	66.2	66.0	62.5	55.6	52.6	45.6	49.7	51.3	57.0	51.7	88.3
SCONE			51.6	69.4	70.0	55.4	55.0	52.1	53.1	66.3	66.0	62.6	55.7	53.0	45.8	49.9	51.4	57.1	52.5	88.2
DUL (Ours)			**51.0**	**69.1**	70.5	55.1	**54.5**	**51.5**	**52.4**	66.2	65.9	62.6	55.7	53.0	45.4	**49.4**	**50.9**	**56.9**	**49.1**	**89.3**

Figure 1: Visualization of different types of uncertainty on semantic OOD test dataset (i.e., Textures) when CIFAR-10 is ID dataset. Without DUL (orange), all three types of uncertainty will increase altogether on OOD. In contrast, DUL (green) increases the distributional uncertainty but decreases the data uncertainty on OOD, which further lead to unchanged overall uncertainty. These observation meets our expectation. We use Eq.17 from [9] to calculate data uncertainty. The distributional uncertainty is shifted by subtracting that on ID dataset.

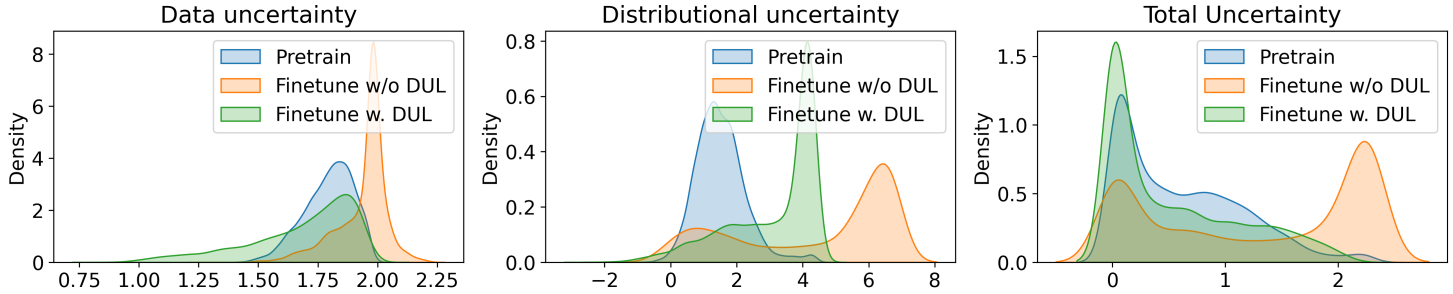

Table 2: Additional results involving new baselines suggested by our reviewers. ID dataset is CIFAR-10. $\mathcal{P}_{\text{train}}^{\text{SEM}}$ is ImageNet-RC. $\mathcal{P}_{\text{test}}^{\text{Cov}}$ is the original CIFAR-10 testset corrupted by Gaussian noise $\mathcal{N}(0, 5)$.

Method	Model generalization		OOD detection	
	ID-Acc ↑	OOD-Acc ↑	FPR ↓	AUC ↑
Entropy	96.04	72.57	6.63	98.72
EBM (finetune)	96.10	79.03	3.61	98.39
POEM	94.32	78.89	3.32	98.99
K+1 Classifier	**96.26**	84.77	3.22	**99.11**
EBM w. DUL	95.19	87.45	6.17	98.28
Entropy w. DUL	96.10	87.41	29.56	95.92
K+1 w. DUL	95.89	**88.59**	7.12	98.35
DUL	96.02	88.01	5.89	98.47
DUL†	96.04	87.53	5.99	98.28
DUL† (100 epochs)	96.15	88.13	**2.45**	98.08

Table 3: We tune the weight of OOD detection regularization term for EBM as well as Entropy and report the FPR (OOD detection metric) and error rate (Err, OOD generalization metric). The experimental settings are the same with Table 2.

Entropy			EBM		
λ	OOD-Err ↓	FPR ↓	λ	OOD-Err ↓	FPR ↓
0	9.55	35.15	0	9.55	20.57
5×10^{-4}	13.58	8.36	1×10^{-4}	9.46	14.69
5×10^{-3}	15.48	6.37	1×10^{-3}	10.32	13.54
5×10^{-2}	17.97	5.71	1×10^{-2}	16.43	8.15
5×10^{-1}	18.53	5.60	1×10^{-1}	24.38	6.11