Appendix

A Regularization-based Calibration Methods

Evidential deep learning (Sensoy et al., 2018) constructs the classification outputs as Dirichlet distributions $Dir(\boldsymbol{\alpha})$ with parameters $\boldsymbol{\alpha} = [\alpha^1, \cdots, \alpha^K]$, and then minimizes the expected distance between the obtained Dirichlet distributions and the labels while regularizing by minimizing the KL divergence between the obtained Dirichlet distributions and the uniform distribution. The loss function is formulated as follows:

$$\sum_{k=1}^{K} y^k \left(\psi(S) - \psi(\alpha^k)\right) + \gamma KL[Dir(\tilde{\boldsymbol{\alpha}}), Dir([1, \cdots, 1])], \tag{7}$$

where $S = \sum_{k=1}^{K} \alpha^k$ is the Dirichlet strengthes, $\psi(\cdot)$ represents the digamma function, γ is a hyperparameter, $\tilde{\boldsymbol{\alpha}} = \boldsymbol{y} + (1 - \boldsymbol{y}) \odot \boldsymbol{\alpha}$, and $Dir([1, \cdots, 1])$ is a uniform Dirichlet distribution.

Penalizing confidence (Pereyra et al., 2017) suggest a confidence penalty term to prevent the deep neural networks from overfitting and producing overconfident predictions. Formally, the loss function of penalizing confidence is defined as:

$$\mathcal{L}_{ce}(f_\theta(\boldsymbol{x}), \boldsymbol{y}) - \gamma \mathcal{H}(f_\theta(\boldsymbol{x})), \tag{8}$$

where γ is a hyperparameter to control the penalizing strength.

B Experimental details

In this section, we present the experimental setup in detail including the backbone model for each dataset (Sec. B.1), descriptions of the datasets (Sec. B.2), evaluation metrics (Sec. B.3), comparison methods (Sec. B.4), comparison experimental results on various metrics (Sec. B.5), and additional experimental results from hyperparameter analysis (Sec. B.6). We are committed to open-sourcing the code related to our research after publication to present more details.

B.1 Backbone model

For the CIFAR-8-2 and CIFAR-80-20 datasets, we use a randomly initialized WideResNet-28-10 (Zagoruyko & Komodakis, 2016) as the backbone network; for the Camelyon17 dataset, we use a DenseNet-121 (Huang et al., 2017) network pre-trained on ImageNet as the backbone network; for other datasets, we use a ResNet-50 (He et al., 2016) network pre-trained on ImageNet as the backbone network.

B.2 Datasets details

The datasets used in the experiments are described in detail here.

- `CIFAR-8-2`: The CIFAR-8-2 dataset is artificially constructed to evaluate the performance of the model when the outlier fraction of the dataset is available. Specifically, we randomly select 8 classes from CIFAR10 (Krizhevsky et al., 2009) as in-distribution samples, while the remaining samples are randomly relabeled as one of the selected classes to serve as outliers. The true outlier fraction of the CIFAR-8-2 dataset is 20%. Since the outlier labels in the cifar-8-2 dataset are randomly generated, the accuracy and ordinal ranking based confidence evaluation metrics lose their meaning on this dataset. Therefore, we do not report these metrics for the outlier dataset in the experimental results.
- `CIFAR-80-20`: Samilar as CIFAR-8-2, we randomly select 80 classes from CIFAR100 (Krizhevsky et al., 2009) dataset, and relabel other samples as one of the selected calsses to serve as outliers. The true outlier fraction of the CIFAR-80-20 dataset is 20%. Since the outlier labels in the cifar-80-20 dataset are randomly generated, the accuracy and ordinal ranking based confidence evaluation metrics lose their meaning on this dataset. Therefore, we do not report these metrics for the outlier dataset in the experimental results.

- `Camelyon17`: Camelyon17 is a pathology image dataset containing over 450,000 lymph node scans from 5 different hospitals, used for detecting cancerous tissues in images (Bandi et al., 2018). Similar to previous work (Koh et al., 2021), we take part of the data from 3 hospitals as the training set. The remaining data from these 3 hospitals, together with data from another hospital, are used as the validation set. The last hospital is used as an outlier test set. Notably, due to differences in pathology staining methods between hospitals, even data within the same hospital can be seen as sampled from multiple subpopulations. We verify on the Camelyon17 dataset whether models can achieve more robust generalization performance when the training set contains multiple subgroups.
- `ImageNetBG`: ImageNetBG is a benchmark dataset for evaluating the dependence of classifiers on image backgrounds (Xiao et al., 2020). It consists of a 9-class subset of ImageNet (ImageNet-9) and provides bounding boxes that allow removing the background. Similar as in previous settings (Yang et al., 2023), we train models on the original IN-9L (with background) set, adjust hyperparameters based on validation accuracy, and evaluate on the test set (in-distribution data), MIXED-RAND, NO-FG and ONLY-FG test set (outlier data).
- `Food101`: Food101 is a commonly used food classification dataset containing 101 food categories with a total of 101,000 images (Bossard et al., 2014). For each category, there are 750 training images and 250 manually verified test images. The training images are intentionally unclean and contain some amount of noise, primarily in the form of intense colors and occasionally wrong labels, which can be seen as outlier data.

B.3 Evaluation metrics

The evaluation metrics used in the experiments are described in detail here.

- `AURC` and `EAURC`: The AURC is defined as the area under the risk-coverage curve (Geifman & El-Yaniv, 2017), where risk represents the error rate and coverage refers to the proportion of samples with confidence estimates exceeding a specified confidence threshold. A lower AURC indicates that correct and incorrect samples can be effectively separated based on the confidence of the samples. However, AURC is influenced by the predictive performance of the model. To allow for meaningful comparisons across models with different performance, Excess-AURC (E-AURC) is proposed by (Geifman et al., 2018) by subtracting the optimal AURC (the minimum possible value for a given model) from the empirical AURC.
- `AUPRErr`: AUPRErr represents the area under the precision-recall curve where misclassified samples (i.e., incorrect predictions) are used as positive examples. This metric can evaluate the capability of the error detector to distinguish between incorrect and correct samples. A higher AUPRErr usually indicates better error detection performance (Corbière et al., 2019).
- `FPR95%TPR`: The FPR95%TPR metric measures the false positive rate (FPR) when the true positive rate (TPR) is fixed at 95%. This metric can be interpreted as the probability that an incorrect prediction is mistakenly categorized as a correct prediction, when the TPR is set to 95%.
- `ECE`: The Expected Calibration Error (ECE) provides a measure of the alignment between the predicted confidence scores and labels. It partitions the confidence sores into multiple equally spaced intervals, computes the difference between accuracy and average confidence in each interval, and then aggregates the results weighted by the number of samples. Lower ECE usually indicates better calibration.
- `NLL`: The Negative Log Likelihood (NLL) measures the log loss between the predicted probabilities and the one-hot label encodings. Lower NLL corresponds to higher likelihood of the predictions fitting the true distribution.
- `Brier`: The Brier score calculates the mean squared error between the predicted probabilities and the one-hot label.

Since the outlier labels in the cifar-8-2 and cifar-80-20 datasets are randomly generated, the accuracy and ordinal ranking based confidence evaluation metrics lose their meaning on this datasets. Therefore, we do not report these metrics for the outlier dataset in the experimental results.

B.4 COMPARISON METHODS

The Comparison methods are described in detail here.

- `ERM` trains the model by minimizing the empirical risk on the training data, using cross-entropy as the loss function.
- `PC` trains the model with cross-entropy loss while regularize the neural networks by penalizing low entropy predictions.
- `LS` is a regularization technique that trainin the neural network with softened target labels.
- `FLSD` refers to a sample-dependent focal loss, where the hyperparameters of the focal loss are set differently for samples with different confidence scores (Mukhoti et al., 2020).
- `FL` refers to focal loss, which implicitly regularize the deep neural network by increasing the weight of samples with lager losses.
- `IFL` conduct a simple modification on the weighting term of original focal loss by assigning larger weights to the samples with larger output confidences.
- `DFL` aims to achieve a better balance between over-confidence and under-confidence by maximizing the gap between the ground truth logit and the highest logit ranked after the ground truth logit.
- `RO`: We conduct ablation studies by removing potential outlier samples during training. Specifically, during training, given B samples, we directly drop the top ηB samples with the highest losses, where η is the predefined outlier fraction.

B.5 ADDITIONAL RESULTS

In Tab. 3 and Tab. 4, we present the experimental results for all evaluation metrics along with the corresponding standard deviations. From the experimental results we can draw similar conclusions as those in the experiments section. Specifically, DRC achieves the best performance 87 times out of 108 experiments, which strongly suggest DRC outperforms previous approaches in calibration.

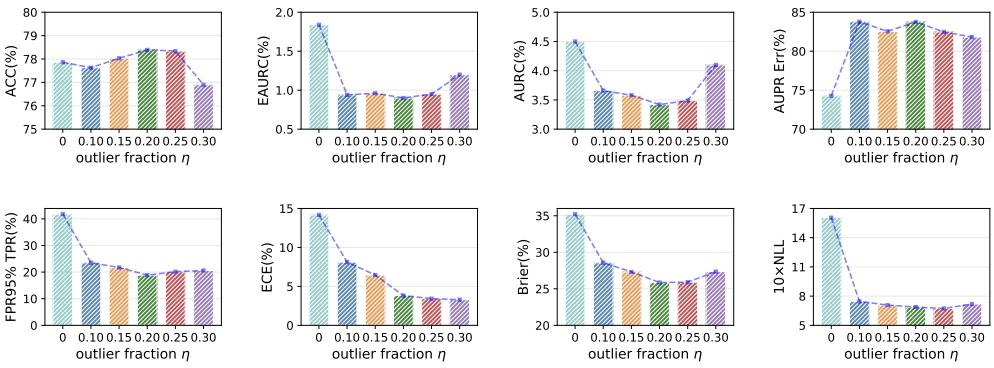

Figure 3: Performance of the model on multiple metrics with varying outlier fraction hyperparameter η, while fixing the regularization strength β to 1, on the CIFAR-8-2 dataset.

B.6 MORE RESULTS OF HYPERPARAMETER ANALYSIS

We present the detailed hyperparameter analysis results on the CIFAR-8-2 and CIFAR-80-20 datasets in Fig.3, Fig.4, Fig.5 and Fig.6. Specifically, to evaluate the effect of the outlier fraction hyperparameter η and regularization strength β on the model, we tune one hyperparameter while fixing the other. From the experimental results we can draw the following conclusions: (1) As shown in Fig.3 and Fig.5, when the set outlier fraction hyperparameter η is close to the true outlier sample ratio, the model can achieve relatively optimal performance. Meanwhile increasing η within a certain range does not significantly degrade the model performance. For example, on most metrics of the CIFAR-8-2 and CIFAR-80-20 datasets, the relatively best performance is achieved at $\eta = 0.2$. (2)

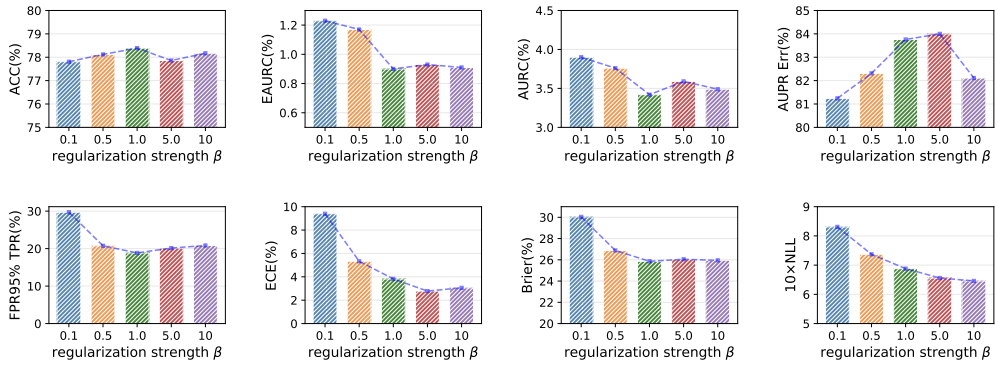

Figure 4: Performance of the model on different metrics with varying regularization strength hyperparameter β, while fixing the outlier fraction hyperparameter η to 0.2, on the CIFAR-8-2 dataset.

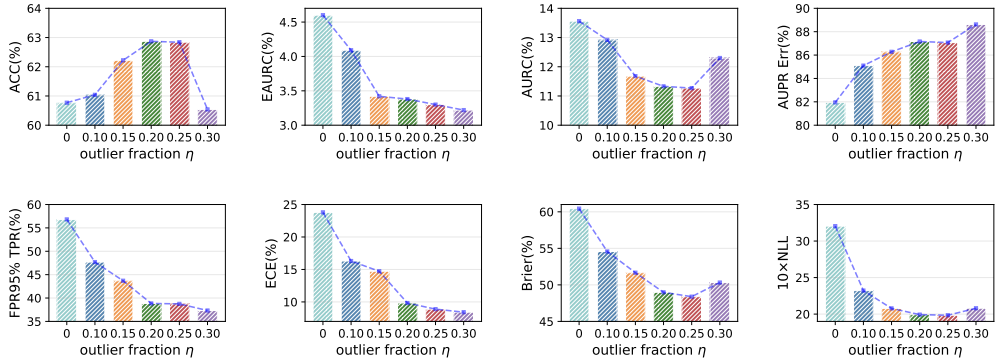

Figure 5: Performance of the model on multiple metrics with varying outlier fraction hyperparameter η, while fixing the regularization strength β to 1, on the CIFAR-80-20 dataset.

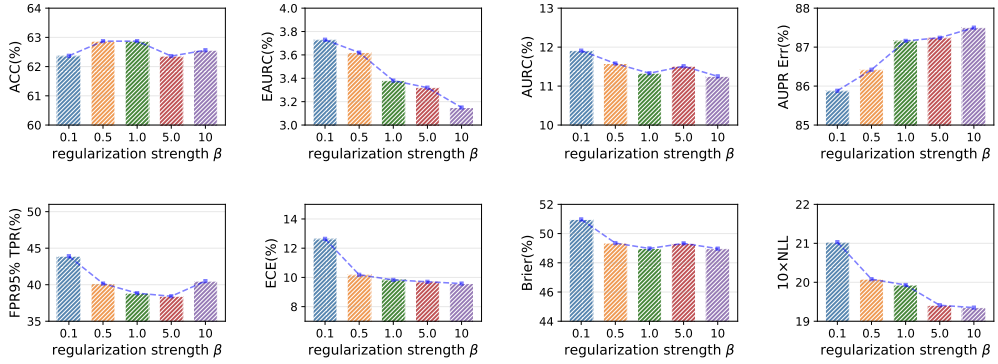

Figure 6: Performance of the model on different metrics with varying regularization strength hyperparameter β, while fixing the outlier fraction hyperparameter η to 0.2, on the CIFAR-80-20 dataset.

As shown in Fig.4 and Fig.6, we can find that increasing the regularization strength β to a certain level yields relatively good performance, after which further increases does not significantly improve the results. For instance, when β exceeds 1, the performance of the model remains relatively stable, with most metrics changing negligibly.

Table 3: The comparison experimental results on different datasets and different methods under the weak augmentation setting. ↓ and ↑ indicate lower and higher values are better respectively. For better presentation, the best and second-best results are in **bold** and underline respectively. For clarity, NLL values are multiplied by 10. Remaining values are reported as percentages (%). For datasets with both ID and outlier test samples, we report the results on all samples and outlier samples. We mark whether the test samples are sampled from ID or outliers in the table. *Compared with other methods,* DRC *achieves excellent performance on different metrics in almost all datasets.*

Dataset	Method	ACC (↑)	EAURC (↓)	AURC (↓)	AUPR Err (↑)	FPR95% TPR(↓)	ECE (↓)	Brier (↓)	NLL (↓)
CIFAR -8-2 (All)	ERM	77.86±0.19	1.84±0.66	4.50±0.71	74.27±5.70	41.81±14.15	14.17±5.27	35.22±5.55	16.06±6.34
	PC	77.36±0.14	2.67±1.08	5.45±1.09	71.12±6.43	48.10±12.39	18.06±2.45	39.34±3.32	23.63±9.21
	LS	77.93±0.23	5.01±2.95	7.65±2.99	74.82±6.05	34.36±9.55	8.37±1.02	30.08±2.61	7.66±0.65
	FLSD	76.79±0.12	2.96±0.16	5.89±0.18	68.88±1.13	56.54±1.52	11.63±0.07	35.42±0.26	11.87±0.15
	FL	77.26±0.25	2.37±0.05	5.18±0.12	70.58±0.18	52.30±0.88	16.31±0.13	37.65±0.30	15.38±0.21
	IFL	78.30±0.18	1.02±0.14	3.57±0.10	82.98±1.27	**18.80±2.05**	5.64±1.04	26.80±0.76	7.92±0.88
	DFL	77.14±0.16	2.35±0.07	5.19±0.05	69.92±1.06	53.53±1.44	16.00±0.09	37.65±0.24	15.48±0.07
	RO	77.57±0.31	4.10±0.44	6.83±0.40	65.33±0.89	57.23±0.75	18.52±0.47	39.97±0.65	18.55±0.60
	DRC	**78.39±0.02**	**0.90±0.04**	**3.42±0.03**	**83.76±0.49**	18.85±1.71	**3.82±0.24**	**25.86±0.21**	**6.88±0.07**
CIFAR -80-20 (All)	ERM	60.77±0.37	4.60±0.18	13.56±0.34	81.95±0.18	56.81±0.30	23.79±0.39	60.44±0.72	32.04±0.66
	PC	62.10±0.28	4.44±0.05	12.75±0.14	81.40±0.31	55.31±1.48	23.77±0.36	59.27±0.53	29.61±0.47
	LS	62.08±0.43	3.90±0.14	12.22±0.33	85.06±0.37	46.26±0.72	**8.31±0.81**	48.99±0.67	**17.79±0.26**
	FLSD	59.44±0.36	5.04±0.14	14.68±0.32	82.15±0.42	55.51±0.96	12.92±0.35	54.81±0.49	23.31±0.17
	FL	60.30±0.12	4.73±0.18	13.92±0.21	81.40±0.17	55.35±0.95	18.30±0.12	56.81±0.26	26.06±0.28
	IFL	61.82±0.46	3.62±0.03	12.07±0.25	86.66±0.31	44.96±0.63	20.80±0.49	56.18±0.62	27.76±0.25
	DFL	59.94±0.21	4.96±0.04	14.34±0.14	81.59±0.37	55.03±0.50	17.24±0.22	56.71±0.40	26.22±0.15
	RO	60.78±0.46	4.96±0.10	13.92±0.19	80.99±0.35	58.50±0.65	24.25±0.85	61.08±0.75	30.90±2.45
	DRC	**62.74±0.44**	**3.19±0.07**	**11.21±0.24**	**87.41±0.21**	**40.02±0.92**	9.53±0.41	**48.76±0.71**	19.31±0.13
Image NetBG (All)	ERM	85.79±0.11	1.23±0.01	2.29±0.01	62.29±0.37	47.97±0.50	4.79±0.41	20.30±0.20	4.71±0.08
	PC	85.57±0.15	1.33±0.07	2.43±0.08	62.42±1.70	49.01±3.92	8.62±0.46	22.45±0.71	6.30±0.44
	LS	**86.62±0.12**	1.34±0.03	2.27±0.05	62.91±0.28	43.95±0.46	10.01±0.30	19.84±0.09	4.92±0.03
	FLSD	85.36±0.16	1.43±0.05	2.56±0.03	61.70±1.63	49.68±2.06	6.30±0.08	21.00±0.04	4.81±0.01
	FL	85.67±0.11	1.29±0.02	2.37±0.03	62.62±0.19	48.03±0.25	1.50±0.11	19.76±0.12	4.42±0.03
	IFL	85.47±0.05	1.26±0.03	2.38±0.03	62.99±0.48	48.07±0.74	6.67±0.32	21.32±0.11	5.21±0.07
	DFL	85.66±0.25	1.42±0.07	2.51±0.11	61.69±0.49	48.87±2.11	1.62±0.07	19.94±0.38	4.50±0.09
	RO	85.95±0.18	1.70±0.03	2.73±0.01	58.32±0.63	52.66±0.28	8.88±0.18	22.56±0.29	6.44±0.08
	DRC	86.12±0.11	**0.89±0.01**	**1.90±0.02**	**67.74±0.34**	**38.18±0.42**	**1.04±0.17**	**18.71±0.18**	**4.41±0.06**
Image NetBG (Outlier)	ERM	81.51±0.15	2.09±0.02	3.91±0.02	62.91±0.36	56.87±0.50	6.32±0.52	26.34±0.30	6.12±0.11
	PC	81.24±0.17	2.23±0.12	4.11±0.14	63.03±1.70	57.12±2.79	11.20±0.57	29.14±0.93	8.18±0.58
	LS	**82.60±0.16**	2.17±0.04	3.78±0.07	63.60±0.31	51.89±0.65	9.87±0.39	25.24±0.10	6.07±0.03
	FLSD	80.94±0.20	2.41±0.07	4.36±0.06	62.41±1.60	57.78±1.66	6.32±0.13	26.88±0.06	6.06±0.02
	FL	81.36±0.13	2.18±0.03	4.04±0.05	63.19±0.16	56.30±0.58	2.08±0.25	25.61±0.15	5.71±0.05
	IFL	81.11±0.04	2.14±0.05	4.05±0.06	63.60±0.43	57.04±0.74	8.69±0.38	27.66±0.13	6.76±0.08
	DFL	81.34±0.35	2.38±0.12	4.24±0.19	62.33±0.46	56.63±1.86	1.75±0.36	25.79±0.53	5.79±0.12
	RO	81.82±0.24	2.73±0.03	4.49±0.03	59.33±0.70	59.97±0.51	11.52±0.22	29.13±0.38	8.33±0.10
	DRC	81.97±0.15	**1.51±0.01**	**3.24±0.04**	**68.38±0.33**	**47.09±0.48**	**1.32±0.22**	**24.17±0.23**	**5.68±0.08**
Food 101 (ID)	ERM	84.99±0.09	1.62±0.01	2.80±0.02	60.33±0.44	52.35±0.68	4.82±0.18	21.90±0.14	5.87±0.03
	PC	85.29±0.16	1.63±0.02	2.77±0.04	59.95±0.32	52.59±0.97	8.12±0.12	22.94±0.29	7.15±0.09
	LS	85.04±0.10	2.21±0.03	3.39±0.04	57.59±0.31	55.80±0.42	10.49±0.07	23.35±0.07	6.79±0.03
	FLSD	85.00±0.07	1.75±0.03	2.93±0.03	58.00±0.42	55.94±1.31	3.54±0.04	21.82±0.04	5.47±0.01
	FL	85.37±0.13	1.68±0.02	2.81±0.03	58.61±0.93	53.37±0.80	1.32±0.10	21.11±0.07	5.32±0.01
	IFL	86.50±0.12	1.52±0.03	2.47±0.05	57.81±0.20	**51.32±0.37**	7.64±0.16	21.29±0.26	6.66±0.12
	DFL	85.50±0.08	1.63±0.02	2.74±0.03	58.52±0.27	52.95±1.04	**0.80±0.15**	20.79±0.10	**5.30±0.01**
	RO	85.42±0.26	1.61±0.03	2.72±0.07	58.67±0.15	53.70±0.06	6.24±0.33	21.93±0.45	6.11±0.14
	DRC	**86.53±0.09**	**1.46±0.02**	**2.41±0.01**	**57.93±0.17**	51.61±0.89	3.66±0.14	**19.95±0.05**	5.44±0.02
Came lyon (Outlier)	ERM	85.75±1.32	3.41±0.66	4.48±0.86	39.42±0.39	73.35±2.54	8.73±1.15	22.56±2.32	4.41±0.54
	PC	85.16±0.47	3.42±0.36	4.58±0.43	40.42±0.61	74.48±1.54	11.44±0.43	25.41±0.93	6.38±0.32
	LS	84.73±1.43	4.41±0.71	5.65±0.94	38.21±0.77	75.98±2.02	13.21±1.57	25.94±1.54	4.29±0.18
	FLSD	86.09±0.73	3.69±0.11	4.71±0.22	37.85±1.08	73.96±0.76	9.12±0.77	21.99±0.60	3.69±0.08
	FL	86.60±0.77	3.36±0.19	4.30±0.27	38.39±1.64	72.46±1.24	3.45±0.77	19.52±0.78	3.23±0.11
	IFL	85.85±0.29	2.92±0.16	3.97±0.13	41.19±1.40	72.31±0.57	11.21±0.32	24.46±0.49	6.31±0.07
	DFL	85.75±0.84	3.03±0.25	4.10±0.36	**41.58±1.13**	71.72±1.09	**2.87±0.25**	19.74±0.92	**3.19±0.13**
	RO	84.36±1.91	4.52±0.64	5.83±0.93	39.43±2.03	75.86±2.24	12.34±2.05	27.14±3.71	7.19±1.32
	DRC	**87.46±1.56**	**2.83±0.43**	**3.66±0.62**	37.96±1.62	**71.18±2.54**	6.39±1.35	**19.34±2.38**	3.52±0.41

Dataset	Method	ECE (↓)	Brier (↓)	NLL (↓)	Dataset	Method	ECE (↓)	Brier (↓)	NLL (↓)
CIFAR -8-2 (Outlier)	ERM	55.69±23.39	136.37±24.91	67.67±29.01	CIFAR -80-20 (Outlier)	ERM	64.26±0.39	151.27±0.57	106.94±1.80
	PC	70.60±10.18	151.27±13.78	99.21±42.13		PC	64.50±1.19	152.05±1.46	95.59±2.37
	LS	36.33±11.02	110.98±9.95	27.55±2.13		LS	29.65±0.92	116.05±0.73	54.92±0.62
	FLSD	57.10±0.44	134.02±0.63	50.15±0.73		FLSD	45.59±0.39	129.95±0.48	76.85±0.56
	FL	67.07±0.65	146.57±0.99	66.37±0.74		FL	53.16±0.35	138.38±0.37	86.54±0.65
	IFL	14.01±4.25	97.20±3.93	26.11±3.04		IFL	46.29±0.38	133.23±0.50	79.58±0.86
	DFL	65.82±0.28	144.78±0.43	66.31±0.46		DFL	52.10±0.26	137.08±0.34	86.46±0.28
	RO	74.56±1.08	156.72±1.46	79.05±2.43		RO	65.14±2.38	152.38±2.96	96.58±8.82
	DRC	**10.24±0.44**	**93.44±0.61**	**23.37±0.41**		DRC	**15.38±0.85**	**109.05±0.49**	**49.97±0.35**

Table 4: The comparison experimental results on different datasets and different methods under the strong augmentation setting. ↓ and ↑ indicate lower and higher values are better respectively. For better presentation, the best and second-best results are in **bold** and underline respectively. For clarity, NLL values are multiplied by 10. Remaining values are reported as percentages (%). For datasets with both ID and outlier test samples, we report the results on all samples and outlier samples. We mark whether the test samples are sampled from ID or outliers in the table. *Compared with other methods,* DRC *achieves excellent performance on different metrics in almost all datasets.*

Dataset	Method	ACC (↑)	EAURC (↓)	AURC (↓)	AUPR Err (↑)	FPR95% TPR(↓)	ECE (↓)	Brier (↓)	NLL (↓)
CIFAR -8-2 (All)	ERM	79.35±0.27	0.70±0.02	2.99±0.07	83.77±0.83	16.29±1.22	6.89±0.78	25.86±0.77	6.78±0.31
	PC	79.65±0.17	0.64±0.03	2.87±0.01	84.54±1.18	12.80±1.53	4.94±0.44	24.35±0.29	6.66±0.20
	LS	79.65±0.10	4.57±1.80	6.80±1.78	75.48±3.08	29.96±5.26	10.01±0.40	28.05±0.89	7.62±0.34
	FLSD	78.22±0.18	2.33±0.06	4.89±0.05	70.65±0.83	50.43±0.94	8.58±0.36	32.14±0.14	10.54±0.11
	FL	78.96±0.09	1.07±0.19	3.46±0.20	78.62±1.84	31.57±6.61	9.51±1.84	29.34±1.73	8.78±1.28
	IFL	79.70±0.09	0.70±0.01	2.92±0.01	**84.86±0.19**	12.85±0.28	4.97±0.34	24.39±0.22	6.82±0.16
	DFL	78.69±0.25	1.97±0.13	4.42±0.07	71.32±1.06	48.43±2.15	13.25±0.24	33.70±0.05	12.47±0.38
	RO	75.38±4.42	2.03±0.47	5.48±1.78	76.19±3.88	43.03±2.23	17.36±1.22	40.15±4.98	23.41±3.95
	DRC	**79.97±0.09**	**0.62±0.02**	**2.78±0.03**	84.09±0.77	**11.50±0.30**	**3.30±0.32**	**23.14±0.09**	**5.98±0.07**
CIFAR -80-20 (All)	ERM	63.68±0.47	3.42±0.51	11.01±0.70	84.17±2.74	47.71±7.28	17.08±4.87	52.00±4.53	25.22±6.82
	PC	64.10±0.52	3.02±0.18	10.42±0.40	86.81±0.14	40.59±0.98	16.09±0.34	50.17±0.79	21.96±0.67
	LS	63.73±0.61	4.17±0.90	11.73±0.63	83.69±3.16	46.24±4.89	**5.07±2.14**	47.08±0.38	18.16±0.75
	FLSD	61.88±0.35	5.02±0.03	13.44±0.16	80.19±0.35	55.50±0.33	16.49±0.06	54.23±0.36	26.09±0.42
	FL	63.06±0.45	3.30±0.07	11.16±0.28	86.46±0.22	41.15±0.98	9.29±0.30	47.72±0.46	18.56±0.30
	IFL	64.43±0.62	3.02±0.09	10.27±0.33	86.66±0.48	41.42±1.10	21.66±1.07	54.25±1.46	29.84±1.70
	DFL	63.47±0.48	3.31±0.10	10.99±0.31	85.89±0.33	41.43±0.72	8.42±1.17	47.08±0.52	18.16±0.35
	RO	60.20±1.04	4.97±0.30	14.23±0.81	81.19±0.12	58.94±0.96	24.73±2.69	62.20±1.83	42.31±3.34
	DRC	**65.90±0.33**	**2.37±0.11**	**8.98±0.24**	**88.38±0.10**	**34.76±0.29**	9.84±0.40	**44.59±0.44**	**17.20±0.19**
Image NetBG (All)	ERM	87.35±0.14	0.98±0.02	1.82±0.04	61.85±1.00	44.06±1.25	3.47±0.10	17.79±0.30	4.04±0.07
	PC	86.88±0.36	1.09±0.06	1.99±0.11	61.90±0.71	45.12±2.37	7.19±0.46	19.95±0.89	5.21±0.35
	LS	**87.54±0.39**	1.26±0.02	2.07±0.07	60.55±0.61	44.85±0.89	4.42±0.27	17.73±0.47	4.20±0.12
	FLSD	86.64±0.34	1.21±0.04	2.15±0.09	61.42±0.94	47.22±0.32	7.51±0.38	19.56±0.33	4.50±0.07
	FL	86.70±0.20	1.17±0.07	2.10±0.08	62.08±1.79	46.23±2.16	3.92±0.13	18.65±0.37	4.21±0.09
	IFL	87.26±0.24	1.03±0.04	1.88±0.06	62.24±0.77	45.32±1.56	4.89±0.27	18.37±0.31	4.32±0.11
	DFL	87.11±0.48	1.17±0.06	2.04±0.13	60.91±1.27	46.51±0.79	2.96±0.26	18.11±0.66	4.09±0.15
	RO	87.21±0.29	1.23±0.05	2.08±0.08	59.35±1.32	48.24±1.78	6.22±0.26	19.22±0.33	4.76±0.09
	DRC	87.53±0.42	**0.94±0.04**	**1.75±0.09**	**62.53±0.60**	**42.47±0.65**	**2.23±0.42**	**17.45±0.51**	**4.01±0.14**
Image NetBG (Outlier)	ERM	83.52±0.15	1.68±0.03	3.12±0.06	62.37±1.06	53.02±1.13	4.61±0.10	23.13±0.37	5.25±0.10
	PC	82.91±0.47	1.83±0.11	3.38±0.19	62.65±0.57	54.36±2.34	9.40±0.63	25.93±1.17	6.77±0.46
	LS	83.81±0.52	2.07±0.03	3.46±0.12	61.05±0.67	53.05±0.90	4.21±0.30	22.86±0.62	5.31±0.16
	FLSD	82.63±0.43	2.05±0.08	3.66±0.16	61.96±0.92	54.95±0.86	7.92±0.44	24.99±0.44	5.67±0.10
	FL	82.72±0.25	1.99±0.11	3.57±0.14	62.59±1.83	54.38±2.00	4.04±0.17	24.01±0.48	5.38±0.12
	IFL	83.41±0.35	1.73±0.06	3.19±0.11	62.87±0.77	53.79±1.17	6.43±0.37	23.87±0.47	5.61±0.16
	DFL	83.28±0.64	1.97±0.10	3.46±0.21	61.44±1.35	54.90±0.88	3.03±0.35	23.36±0.86	5.24±0.21
	RO	83.41±0.37	2.04±0.09	3.51±0.14	59.99±1.34	56.03±1.06	8.14±0.34	24.93±0.41	6.18±0.12
	DRC	**83.83±0.55**	**1.58±0.05**	**2.97±0.15**	**63.01±0.62**	**51.31±0.14**	**2.97±0.56**	**22.59±0.67**	**5.18±0.18**
Food 101 (ID)	ERM	87.26±0.14	1.25±0.02	2.10±0.02	58.21±1.04	50.47±1.13	2.37±0.08	18.44±0.12	4.68±0.04
	PC	87.54±0.08	1.24±0.01	2.05±0.02	56.87±0.23	50.58±0.30	5.87±0.12	19.17±0.11	5.45±0.02
	LS	87.33±0.03	1.80±0.03	2.64±0.03	53.76±0.22	54.59±0.56	19.70±0.16	23.58±0.03	6.91±0.01
	FLSD	85.98±0.09	1.55±0.01	2.58±0.01	57.34±0.55	54.40±0.37	6.04±0.04	20.90±0.08	5.21±0.02
	FL	86.32±0.22	1.45±0.02	2.43±0.05	58.26±0.43	51.60±0.45	3.62±0.12	19.91±0.17	4.95±0.04
	IFL	**87.67±0.07**	**1.22±0.04**	**2.01±0.03**	57.87±0.80	**49.71±1.12**	5.05±0.07	18.58±0.07	5.10±0.01
	DFL	86.80±0.11	1.40±0.04	2.31±0.06	57.31±0.47	52.08±1.04	**1.29±0.13**	19.15±0.23	4.80±0.06
	RO	87.23±0.01	1.26±0.03	2.12±0.03	57.80±1.23	50.83±1.60	3.04±0.08	18.65±0.12	4.79±0.05
	DRC	87.32±0.12	1.25±0.01	2.09±0.03	**58.27±0.42**	50.33±0.76	2.17±0.13	**18.34±0.13**	**4.67±0.02**
Came lyon (Outlier)	ERM	90.21±0.38	1.41±0.06	1.91±0.10	38.69±0.55	63.63±0.88	4.60±0.19	14.89±0.51	2.59±0.08
	PC	87.30±0.40	1.97±0.10	2.81±0.14	**40.47±0.75**	70.33±0.54	9.91±0.80	21.92±1.14	5.09±0.54
	LS	92.52±0.48	1.28±0.16	1.57±0.19	34.84±0.71	58.96±1.14	18.69±0.36	18.66±0.47	3.49±0.06
	FLSD	92.24±0.28	1.12±0.05	1.43±0.07	35.29±0.31	59.69±1.19	12.81±0.27	15.83±0.17	2.85±0.03
	FL	92.30±0.24	1.11±0.04	1.41±0.06	35.17±0.12	59.67±1.09	12.91±0.28	15.82±0.10	2.85±0.02
	IFL	88.99±0.85	1.59±0.19	2.23±0.29	40.02±0.69	65.91±1.59	7.15±0.92	17.71±1.57	3.36±0.33
	DFL	90.60±1.30	1.72±0.21	2.19±0.34	36.44±2.74	61.35±0.52	7.18±0.96	15.39±0.93	2.66±0.11
	RO	91.55±0.58	2.62±0.33	2.99±0.38	31.57±1.27	64.19±2.47	7.10±0.60	15.28±1.16	5.31±0.58
	DRC	**93.43±0.22**	**0.83±0.06**	**1.05±0.07**	34.34±0.92	**57.07±1.58**	**2.30±0.38**	**10.16±0.39**	**1.88±0.10**

Dataset	Method	ECE (↓)	Brier (↓)	NLL (↓)	Dataset	Method	ECE (↓)	Brier (↓)	NLL (↓)
CIFAR -8-2 (Outlier)	ERM	26.42±2.31	103.12±1.77	27.70±1.06	CIFAR -80-20 (Outlier)	ERM	47.87±13.87	134.26±14.91	87.04±27.59
	PC	15.17±1.88	96.69±1.59	25.52±0.86		PC	37.96±0.75	124.76±0.48	67.83±1.09
	LS	39.34±4.34	113.79±4.10	29.97±1.45		LS	30.26±3.68	116.36±2.94	58.00±2.39
	FLSD	52.43±0.34	127.95±0.66	45.84±0.71		FLSD	55.18±0.87	140.49±0.93	94.46±1.79
	FL	43.23±7.89	118.76±8.28	37.93±6.23		FL	30.87±1.14	116.85±0.95	60.91±0.90
	IFL	14.40±1.39	96.78±0.89	25.34±0.60		IFL	54.86±2.30	141.27±2.32	93.67±4.71
	DFL	60.51±1.02	138.24±1.58	55.67±2.37		DFL	29.68±2.15	116.09±1.92	59.39±1.78
	RO	58.97±12.40	137.18±14.39	74.23±24.88		RO	65.26±6.81	152.99±7.96	128.48±14.80
	DRC	**8.81±1.51**	**92.69±0.66**	**22.93±0.35**		DRC	**20.57±0.67**	**112.30±0.61**	**51.92±0.33**

C PROOF OF THEOREM 1

Following Bai et al. (2021), we have

$$p - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = p) = p - \mathbb{E}_Z[\sigma(\frac{\|\boldsymbol{w}^*\|}{\|\hat{\boldsymbol{w}}\|} \cos\hat{\theta} \cdot \sigma^{-1}(p)) + \sin\hat{\theta} \cdot \|\boldsymbol{w}^*\| Z],$$

where $\cos\hat{\theta} = \frac{\hat{\boldsymbol{w}}^\top \boldsymbol{w}^*}{\|\hat{\boldsymbol{w}}\| \cdot \|\boldsymbol{w}^*\|}$.

We first compute the calibration error for the baseline method

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{w}^\top \boldsymbol{x}_i - \tilde{y}_i)^2,$$

where $\tilde{y}_i = (1-\epsilon) y_i + \epsilon/2$. As $d/n = o(1)$, we have

$$\hat{\boldsymbol{w}} = (\mathbb{E}[x_i x_i^\top])^{-1} \mathbb{E}[x_i \tilde{y}_i] + o(1) = \frac{1-\epsilon}{1 + \|\boldsymbol{w}^*\|^2} \mathbb{E}[x_i y_i] + o(1) = (1-\epsilon)(1-2\eta) \cdot \boldsymbol{w}^* + o(1).$$

In the above derivation, the first equation uses Sherman–Morrison formula.

As a result, we have $\cos\hat{\theta} = 1 - o(1)$ as n grows, and therefore when $n \to \infty$,

$$p - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_{baseline}(X) = p) = p - \sigma(\frac{1}{(1-\epsilon)(1-2\eta)} \sigma^{-1}(p)).$$

Now, for the DRC, when the initialization parameter $\boldsymbol{\theta}^{(0)}$ satisfies $\|\boldsymbol{\theta}^{(0)} - \boldsymbol{w}^*\| \le c_1$ for a sufficiently small constant $c_1 > 0$, there will be only $o(1)$ outliers left, and therefore

$$p - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_{\text{DRC}}(X) = p) = p - \sigma(\frac{1}{1-\eta} \sigma^{-1}(p)).$$

By the monotonicity of σ and the nonnegativity of $\sigma^{-1}(p)$ when $p > 1/2$. We have

$$|p - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_{\text{DRC}}(X) = p)| < |p - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_{baseline}(X) = p)|.$$

Taking the expectation of p for both sides, we have

$$\textit{ECE}[\hat{\mathbb{P}}_{\text{DRC}}] < \textit{ECE}[\hat{\mathbb{P}}_{baseline}].$$