Supplementary Materials: Confidence-Aware Learning for Deep Neural Networks

S1. Experimental Details: Ordinal Ranking

S1.1. Evaluation Metrics

AURC & E-AURC AURC measures the area under the curve drawn by plotting the risk according to coverage. The coverage indicates the ratio of samples whose confidence estimates are higher than some confidence threshold, and the risk, also known as the selective risk (Geifman & El-Yaniv, 2017), is an error rate computed by using those samples. A low value of AURC implies that correct and incorrect predictions can be well-separable by confidence estimates associated with samples.

Inherently, AURC is affected by the predictive performance of a model. To have a unitless performance measure that can be applied across models, Geifman et al. (2019) introduce a normalized AURC, named Excess-AURC (E-AURC). E-AURC can be computed by subtracting the optimal AURC, the lowest possible value for a given model, from the empirical AURC. For a detailed description, please refer to Geifman et al. (2019).

AUPR-Error AUPR measures the area under the precision-recall curve. The precision-recall curve is a graph showing the precision = TP/(TP+FP) against recall = TP/(TP+FN), where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The AUPR-ERROR represents the area under precision-recall curve where misclassified samples (i.e., errors) are used as positives. This is used as the primary metric to evaluate the failure prediction performance in Corbière et al. (2019).

FPR-at-95%-TPR FPR-at-95%-TPR measures the false positive rate (FPR) = FP/(FP+TN) when the true positive rate (TPR) = TP/(TP+FN) is 95%, where TP, TN, FP, and FN denotes true positives, true negatives, false positives, and false negatives, respectively. It can be interpreted as the probability that an example predicted incorrectly is misclassified as a correct prediction when TPR is equal to 95%.

ECE Expected calibration error (ECE) (Naeini et al., 2015) is a metric that approximates the difference in expectation between accuracy and confidence. As an approximation, ECE partitions the probability interval into a fixed number of bins. Then, each bin B_m has an interval $(\frac{m-1}{M}, \frac{m}{M}], m = 1, ..., M$ where M is the number of bins. With these bins, ECE can be computed as

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|$$

where n is the total number of samples, acc(B_m) denotes the accuracy computed from samples in B_m, and conf(B_m) is the average confidence scores of samples in B_m.

NLL Negative log likelihood (NLL) is a standard measure for evaluating the quality of predictive probability, which is computed as

$$\text{NLL} = -\sum_{i=1}^{n} \log P\left(y = y_i|\mathbf{x}_i, \mathbf{w}\right).$$

Brier Score Brier score (Brier, 1950) can be interpreted as the average mean squared error between the predicted probability and one-hot encoded label. It can be computed as

$$\text{Brier} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(P\left(y = k|\mathbf{x}_i, \mathbf{w}\right) - t_k\right)^2$$

where $t_k = 1$ if $k = y_i$, and 0 otherwise.

S1.2. Experimental Settings

Datasets CIFAR-10 and CIFAR-100 are the datasets for a multi-class image classification task. They consist of 50K training images and 10K test images of size 32×32 with 10 and 100 classes, respectively. The Street View House Numbers (SVHN) dataset (Netzer et al., 2011) contains 73,257 training images and 26,032 test images of size 32×32 with 10 classes of digits.

MCdropout VGG-16 for MCdropout is the one used in Geifman et al. (2019).[1] Specifically, a dropout layer with a dropout rate $p = 0.3$ is added after the first convolutional layer, and dropout layers with $p = 0.4$ are applied to other convolutional layers except ones followed by a max pooling layer. For fully connected layers, dropout with $p = 0.5$ is used. PreAct-ResNet110 for MCdropout comes from Zhang et al. (2019). Dropout layers with $p = 0.2$ are applied to all convolutional layer, and a dropout layer with $p = 0.1$

[1] https://github.com/geifmany/uncertainty_ICLR

is added before the last fully connected layer. Note that this architecture from Zhang et al. (2019) was determined through the validation process. DenseNet-BC already has dropout layers and we set the dropout rate of them to 0.2 as used in the original paper (Huang et al., 2017). In the experiments, we compute 50 stochastic predictions and the entropy on the average predicted probabilities is used as an uncertainty estimate.

Aleatoric+MCdropout To consider aleatoric uncertainty, a Gaussian distribution whose mean is a model's prediction is placed over the logit space as proposed in Kendall & Gal (2017). The models for Aleatoric+MCdropout are the same as used for MCdropout except that an additional output layer is attached to produce the variance of the Gaussian distribution. With this Gaussian distribution, 50 logit vectors are sampled and averaged to compute a cross-entropy loss during training. Like MCdropout, we use 50 stochastic predictions and the entropy is used to estimate uncertainty.

AES Average early stopping (AES) is a snapshot ensemble approach motivated by the observation that easy samples are learned earlier during training while hard samples are not. To leverage this for confidence estimation, AES method provides the average confidence estimates from the ensemble of model snapshots. Geifman et al. (2019) suggests an ensemble with k models at epochs $i \in F$ where F is a set of k evenly spaced integers between $0.4T$ and T. Here, T denotes the total number of epochs. In the experiments, we consider $k = 10$ and $k = 30$.

S1.3. Results

Table S1, S2 and S3 shows the complete experimental results to evaluate ordinal ranking performance on CIFAR-10, CIFAR-100 and SVHN, respectively. For CRL models, we consider the maximum class probability (CRL-softmax), negative entropy (CRL-entropy), and margin (CRL-margin) as a confidence function, respectively. Regardless of the confidence function, it is observed that CRL improves the quality of confidence estimates. Compared to other methods that require multiple predictions, CRL models consistently yield comparable or better performance.

Figure S1 shows the risk-coverage (RC) curve plots from PreAct-ResNet110 on CIFAR-10/100 and SVHN datasets. A score in parentheses is the AURC value associated with each model. For this figure, the model that shows the median performance among five repeated runs is selected.

Tables S4 and S5 show ordinal ranking performance of CRL ensembles with $\lambda = 0.5$ and $\lambda = 1.0$, respectively.

S2. Experimental Details: Out-of-Distribution Detection

S2.1. Evaluation metrics

Detection Error Detection error measures the minimum possible error rate over all possible thresholds when separating in- and out-of-distribution samples.

AUROC The area under the receiver operating characteristic curve (AUROC) measures the area under the curve drawn by plotting the true positive rate against the false positive rate.

AUPR-In & AUPR-Out AUPR measures the area under the precision-recall curve. AUPR-In and AUPR-Out use in- and out-of-distribution samples as positives, respectively.

S2.2. Experimental Settings

Datasets The TinyImageNet is a subset of ImageNet dataset that contains 10,000 test images with 200 classes. The LSUN dataset consists of 10,000 images of 10 different scenes (Yu et al., 2015). The iSUN dataset is a subset of LSUN images and consists of 8,925 images. These datasets are used as out-of-distribution datasets, and all images are resized to 32×32.

ODIN ODIN (Out-of-DIstribution detector for Neural networks) (Liang et al., 2018) is a simple and effective post-processing method for out-of-distribution detection. ODIN consists of two steps: temperature scaling and adding small perturbations to inputs. Through a manipulation of temperature constant T, the softmax scores of in- and out-of-distribution images can be distinguishable by pusing them further apart from each other. In addition, an input is pre-processed by adding small perturbations to decrease the softmax score. The perturbations can be computed as the gradient of loss with respect to the input, and they are added to the input with a multiplicative constant ϵ. To find the hyperparameters T and ϵ, a small hold-out set from out-of-distribution dataset was used following to the procedure in the original paper.

Mahalanobis Lee et al. (2018) proposed the Mahalanobis distance-based confidence score to identify out-of-distribution samples from the finding that the trained features of deep neural networks follow the class-conditional Gaussian distribution. To further enhance the detection performance, it adds small perturbations ϵ to an input similar to ODIN, and combines the confidence scores from all layers in a deep neural network. Concretely, the scores are computed by weighted averaging and these weights are determined by training a logistic regression model using a validation dataset. The optimal value of ϵ was chosen via validation process as described in the original paper.

S2.3. Results

Table S6 shows full out-of-distribution detection results including those from iSUN dataset. Since iSUN is a subset of LSUN, the detection performances on iSUN are similar to those on LSUN.

S3. Experimental Details: Active Learning

S3.1. Experimental Settings

Since query strategies for active learning are based on uncertainty, there exists a risk that samples selected to be labeled are overlapped, i.e., they might have redundant information. To avoid this issue, we select the samples from a random subset of the unlabeled pool $\mathcal{D}_U^S$ at S-th stage. We set the size of subset to 10,000. Beluch et al. (2018) and Yoo & Kweon (2019) are also used this simple scheme to address the redundancy issue.

The proposed method requires counting correct prediction events of all training samples. Hence, incremental learning with newly labeled samples cannot be applied to CRL models. For a fair comparison, we initialize all models including comparison targets at the beginning of every stage, i.e., all models are trained from scratch with their labeled dataset. To control the unexpected effect of random initialization, all models share the same random seed at each stage.

Query Strategy We consider the following query strategies (i.e., sampling methods) for comparison: random sampling, entropy-based sampling, core-set sampling (Sener & Savarese, 2018), and entropy-based sampling with MC-dropout. Random sampling is selecting samples to be labeled randomly. Entropy-based sampling selects samples whose entropy of predicted class probability is high. Entropy-based sampling with MCdropout differs from just entropy-based sampling in that it measures entropy on the average predicted class probabilities obtained by 50 stochastic predictions. Core-set sampling focuses on the representativeness of samples, which can be implemented by K-Center-Greedy algorithm. Following to Sener & Savarese (2018), we use the l_2 distance between activations of the last fully connected layer to measure the diversity of samples.

S3.2. Results

Table S7 shows the classification accuracy values of sampling strategies at each active learning stage.

Table S1. Comparison of the quality of confidence estimates on CIFAR-10. The means and standard deviations over five runs are reported. ↓ and ↑ indicate that lower and higher values are better respectively. AURC and E-AURC values are multiplied by 10^3, and NLL are multiplied by 10 for clarity. All remaining values are percentage. **Red** and **blue** represent the best performance among single models and the methods requiring multiple predictions, respectively.

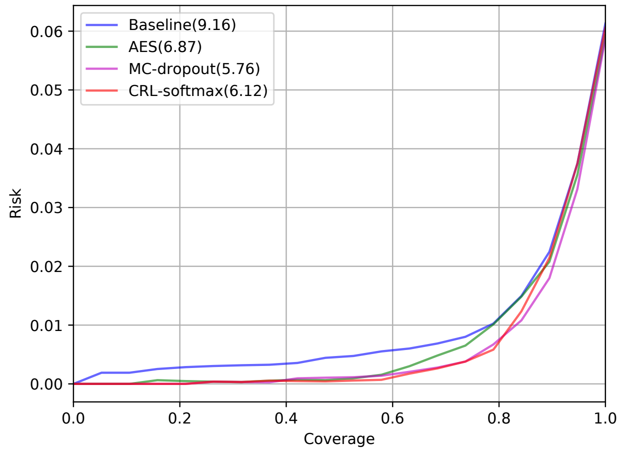
Dataset Model	Method	ACC (↑)	AURC (↓)	E-AURC (↓)	AUPR-Err (↑)	FPR-95% TPR (↓)	ECE (↓)	NLL (↓)	Brier (↓)
	Baseline	93.74±0.14	7.10±0.31	5.10±0.26	44.19±0.34	41.43±0.38	5.20±0.11	3.79±0.11	11.30±0.21
	CRL-entropy	93.84±0.12	**6.77±0.16**	4.83±0.16	46.16±2.87	41.35±3.03	2.47±0.19	2.47±0.03	9.99±0.09
	CRL-softmax	93.82±0.18	6.78±0.18	**4.83±0.08**	**46.79±1.75**	**40.21±2.18**	**1.24±0.20**	**2.09±0.04**	**9.33±0.21**
CIFAR-10	CRL-margin	**93.88±0.12**	7.13±0.23	5.21±0.16	43.26±1.79	44.20±0.94	1.55±0.13	2.73±0.07	9.81±0.16
VGG-16	MCdropout	93.78±0.27	6.72±0.28	4.72±0.19	45.08±2.14	41.52±2.83	1.11±0.19	1.93±0.05	9.34±0.39
	Aleatoric+MC	93.91±0.13	6.57±0.29	4.68±0.22	44.67±1.76	41.68±1.86	**0.86±0.12**	**1.89±0.05**	**9.08±0.24**
	AES(k=10)	**93.97±0.12**	7.15±0.25	5.30±0.25	44.47±1.00	41.01±1.75	1.61±0.27	2.06±0.04	9.26±0.15
	AES(k=30)	93.96±0.17	**6.50±0.10**	**4.64±0.09**	**45.36±3.02**	**38.60±1.51**	1.82±0.25	1.95±0.03	9.23±0.15
	Baseline	94.11±0.20	9.11±0.44	7.34±0.39	42.70±1.59	40.42±2.30	4.46±0.16	3.34±0.13	10.19±0.32
	CRL-entropy	**94.24±0.11**	**6.01±0.18**	4.33±0.13	43.15±0.43	41.65±2.66	**0.79±0.12**	1.97±0.02	**8.74±0.12**
	CRL-softmax	94.00±0.12	6.02±0.26	**4.21±0.19**	45.20±1.15	**38.81±1.59**	1.23±0.18	**1.81±0.04**	8.85±0.20
CIFAR-10	CRL-margin	93.83±0.10	6.28±0.13	4.34±0.07	**45.46±1.07**	39.92±1.27	1.12±0.16	1.87±0.01	9.07±0.09
ResNet110	MCdropout	94.25±0.00	**5.48±0.19**	**3.80±0.16**	45.21±2.19	36.74±3.06	1.45±0.15	1.88±0.05	8.48±0.13
	Aleatoric+MC	**94.33±0.09**	6.02±0.33	4.38±0.30	45.55±0.87	38.72±1.82	**1.25±0.07**	**1.80±0.03**	**8.36±0.12**
	AES(k=10)	94.22±0.22	6.71±0.54	5.00±0.44	44.31±2.00	39.80±2.35	1.38±0.15	1.94±0.05	8.82±0.32
	AES(k=30)	94.20±0.23	5.80±0.28	4.09±0.25	**47.15±1.93**	**36.37±2.85**	1.61±0.20	1.82±0.04	8.69±0.29
	Baseline	94.87±0.23	5.15±0.35	3.82±0.30	44.21±2.21	36.35±2.02	3.20±0.20	2.23±0.09	8.33±0.37
	CRL-entropy	**94.98±0.15**	4.95±0.30	3.67±0.26	40.67±1.50	42.12±2.06	**0.69±0.15**	1.66±0.03	**7.67±0.19**
	CRL-softmax	94.71±0.09	**4.92±0.14**	**3.49±0.94**	45.16±2.12	**36.13±3.35**	0.87±0.07	**1.60±0.02**	7.84±0.17
CIFAR-10	CRL-margin	94.42±0.19	5.26±0.23	3.68±0.18	**45.36±3.22**	36.67±2.19	0.95±0.11	1.65±0.03	8.17±0.18
DenseNet	MCdropout	94.69±0.25	5.30±0.38	3.85±0.28	45.64±2.65	36.61±2.38	1.20±0.09	1.73±0.05	7.92±0.28
	Aleatoric+MC	94.73±0.19	5.17±0.20	3.76±0.14	**45.67±3.18**	34.69±1.03	1.25±0.06	1.72±0.04	7.80±0.16
	AES(k=10)	**95.00±0.14**	5.31±0.32	4.04±0.26	43.29±1.83	37.13±2.69	**1.00±0.10**	1.66±0.04	7.65±0.27
	AES(k=30)	94.99±0.18	**4.70±0.20**	**3.43±0.15**	45.39±2.02	**34.37±1.70**	1.18±0.09	**1.58±0.04**	**7.57±0.26**

Table S2. Comparison of the quality of confidence estimates on CIFAR-100. The means and standard deviations over five runs are reported. ↓ and ↑ indicate that lower and higher values are better respectively. AURC and E-AURC values are multiplied by 10^3, and NLL are multiplied by 10 for clarity. All remaining values are percentage. **Red** and **blue** represent the best performance among single models and the methods requiring multiple predictions, respectively.

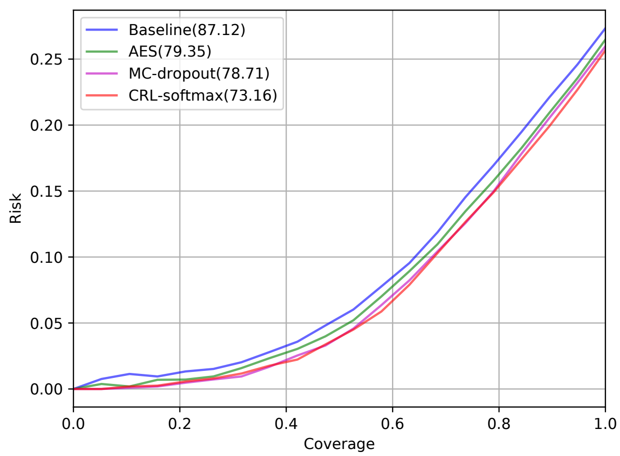
Dataset Model	Method	ACC (↑)	AURC (↓)	E-AURC (↓)	AUPR-Err (↑)	FPR-95% TPR (↓)	ECE (↓)	NLL (↓)	Brier (↓)
	Baseline	73.49±0.34	77.33±1.15	38.61±0.66	68.59±0.64	62.01±0.39	19.81±0.33	17.77±0.37	44.85±0.51
	CRL-entropy	**74.71±0.19**	**70.19±1.53**	35.11±1.13	68.70±1.08	**59.15±2.19**	**11.62±0.32**	**12.42±0.10**	**38.16±0.39**
	CRL-softmax	74.06±0.18	71.83±0.47	**34.84±0.57**	**69.60±1.11**	59.47±1.01	13.86±0.27	13.10±0.12	39.42±0.19
CIFAR-100	CRL-margin	74.06±0.27	75.91±0.76	38.93±0.74	67.59±1.04	59.74±1.62	12.16±0.24	13.67±0.16	38.79±0.36
VGG-16	MCdropout	73.06±0.42	77.36±1.15	37.85±0.51	67.68±0.95	62.39±2.16	3.37±0.37	10.05±0.02	36.59±0.29
	Aleatoric+MC	73.12±0.28	77.31±1.00	37.43±0.42	67.67±0.53	63.53±0.81	**3.22±0.19**	10.02±0.04	36.63±0.21
	AES(k=10)	74.68±0.25	72.25±1.13	37.09±0.58	67.69±0.76	**60.88±0.92**	7.42±0.26	10.02±0.11	35.83±0.36
	AES(k=30)	**74.78±0.30**	**68.99±1.24**	**34.13±0.74**	**67.72±0.95**	61.20±1.40	7.85±0.30	**9.64±0.19**	**35.64±0.38**
	Baseline	72.85±0.30	87.24±1.21	46.50±1.09	66.01±0.43	66.03±1.52	16.58±0.16	15.09±0.14	42.83±0.38
	CRL-entropy	73.73±0.38	75.77±1.81	37.78±1.01	**67.62±1.32**	**61.83±1.46**	**10.37±0.40**	11.23±0.15	38.03±0.53
	CRL-softmax	74.16±0.32	73.59±1.39	**36.90±1.08**	67.23±1.13	62.56±1.26	11.52±0.36	10.87±0.05	37.71±0.44
CIFAR-100	CRL-margin	**74.66±0.13**	**73.26±0.30**	38.04±0.56	63.27±0.59	66.64±1.33	10.77±0.21	**10.50±0.12**	**36.93±0.23**
ResNet110	MCdropout	74.08±0.00	75.47±1.07	38.53±1.13	66.14±1.68	64.59±1.46	5.35±0.32	10.06±0.15	36.06±0.38
	Aleatoric+MC	**74.50±0.24**	**73.26±0.83**	**37.56±0.95**	65.65±0.91	**63.53±1.78**	**2.68±0.25**	**9.24±0.13**	**34.96±0.20**
	AES(k=10)	73.65±0.29	79.12±1.07	40.88±0.49	66.72±0.74	63.81±1.40	8.90±0.15	10.67±0.13	37.67±0.37
	AES(k=30)	73.67±0.32	76.69±1.32	38.52±0.96	**67.13±0.76**	64.23±0.95	9.33±0.20	10.17±0.11	37.61±0.39
	Baseline	75.39±0.29	71.75±0.89	38.63±0.72	65.18±1.71	63.30±1.93	12.67±0.25	11.54±0.08	37.26±0.21
	CRL-entropy	76.24±0.28	64.33±1.19	33.56±0.54	**65.36±0.28**	**61.36±0.92**	**8.02±0.39**	9.60±0.09	34.04±0.44
	CRL-softmax	76.82±0.26	61.77±1.07	**32.57±0.81**	65.22±1.40	61.79±2.20	8.59±0.17	9.11±0.09	33.39±0.28
CIFAR-100	CRL-margin	**77.09±0.18**	**61.51±0.99**	33.00±0.65	61.73±0.64	64.23±1.35	8.42±0.17	**8.97±0.10**	**33.06±0.28**
DenseNet	MCdropout	75.80±0.36	66.92±1.45	34.97±0.46	65.11±1.10	63.27±1.47	**5.59±0.33**	9.42±0.14	34.02±0.38
	Aleatoric+MC	75.50±0.39	67.87±1.55	35.05±0.65	65.92±1.38	**61.69±1.79**	6.01±0.22	9.45±0.13	34.25±0.47
	AES(k=10)	**76.10±0.16**	67.18±0.37	36.04±0.18	64.82±0.83	62.59±0.69	6.78±0.37	9.39±0.04	34.04±0.14
	AES(k=30)	76.05±0.12	**65.22±0.73**	**33.95±0.68**	**65.94±0.84**	62.17±0.54	7.38±0.22	**9.04±0.04**	**33.96±0.16**

Table S3. Comparison of the quality of confidence estimates on SVHN. The means and standard deviations over five runs are reported. ↓ and ↑ indicate that lower and higher values are better respectively. AURC and E-AURC values are multiplied by 10^3, and NLL are multiplied by 10 for clarity. All remaining values are percentage. **Red** and **blue** represent the best performance among single models and the methods requiring multiple predictions, respectively.

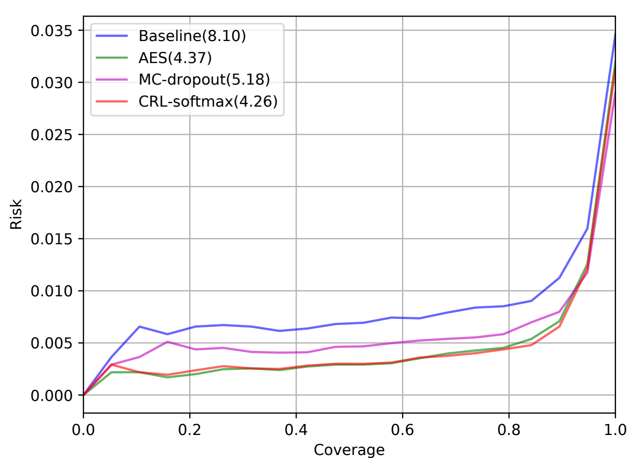
Dataset Model	Method	ACC (↑)	AURC (↓)	E-AURC (↓)	AUPR-Err. (↑)	FPR-95% TPR (↓)	ECE (↓)	NLL (↓)	Brier (↓)
	Baseline	96.20±0.10	5.97±0.28	5.24±0.28	41.15±0.95	32.08±0.56	3.15±0.11	2.69±0.05	6.86±0.17
	CRL-entropy	96.55±0.10	**4.31±0.10**	**3.72±0.10**	**44.39±2.87**	**28.34±1.07**	1.15±0.10	1.55±0.03	5.54±0.11
	CRL-softmax	**96.55±0.07**	4.47±0.10	3.86±0.08	42.82±1.35	29.82±1.42	**0.88±0.12**	**1.52±0.03**	**5.44±0.10**
SVHN	CRL-margin	96.49±0.05	4.50±0.15	3.88±0.13	42.19±0.60	29.18±0.66	0.95±0.03	1.86±0.02	5.67±0.10
VGG-16	MCdropout	96.79±0.05	4.64±0.34	4.12±0.31	41.62±1.21	27.46±0.95	**0.36±0.02**	**1.25±0.03**	**4.96±0.11**
	Aleatoric+MC	**96.80±0.01**	4.86±0.26	4.34±0.26	41.14±0.60	27.60±1.45	0.38±0.07	1.26±0.01	4.99±0.02
	AES(k=10)	96.54±0.09	4.59±0.10	3.98±0.11	43.48±0.86	27.40±0.99	0.54±0.09	1.34±0.01	5.31±0.06
	AES(k=30)	96.58±0.08	**4.27±0.14**	**3.69±0.12**	**43.53±1.16**	**25.20±1.47**	0.50±0.04	1.28±0.01	5.21±0.08
	Baseline	96.45±0.06	8.02±0.76	7.38±0.75	38.83±1.79	35.78±1.45	2.79±0.06	2.38±0.04	6.25±0.12
	CRL-entropy	96.80±0.01	4.12±0.06	3.60±0.06	41.18±1.89	27.81±0.77	1.13±0.05	1.37±0.01	5.12±0.03
	CRL-softmax	96.81±0.09	4.25±0.12	3.74±0.14	**43.46±1.78**	27.71±0.56	**0.85±0.09**	**1.31±0.02**	4.97±0.12
SVHN	CRL-margin	**96.83±0.09**	**4.09±0.14**	**3.58±0.15**	42.32±2.42	**27.00±1.27**	0.86±0.06	1.36±0.02	**4.93±0.08**
ResNet110	MCdropout	97.00±0.00	4.99±0.35	4.53±0.34	39.10±0.94	28.69±2.22	0.65±0.07	1.29±0.01	4.73±0.13
	Aleatoric+MC	**97.01±0.04**	5.54±0.24	5.09±0.23	38.71±1.08	31.60±0.50	0.54±0.05	1.25±0.01	**4.69±0.05**
	AES(k=10)	96.77±0.05	4.41±0.17	3.89±0.16	43.56±2.51	27.39±1.34	0.43±0.11	1.26±0.01	4.97±0.05
	AES(k=30)	96.81±0.05	**4.23±0.13**	**3.72±0.14**	**43.64±1.48**	**26.09±1.54**	**0.33±0.03**	**1.21±0.02**	4.89±0.05
	Baseline	96.40±0.08	7.70±0.41	7.00±0.39	39.43±0.78	34.23±1.21	2.51±0.07	2.10±0.05	6.13±0.15
	CRL-entropy	**96.68±0.07**	**4.27±0.34**	**3.72±0.33**	42.08±2.15	28.76±1.58	0.84±0.05	1.37±0.02	5.20±0.08
	CRL-softmax	96.61±0.12	4.47±0.14	3.89±0.13	**43.35±0.81**	28.35±1.62	0.85±0.06	1.38±0.04	5.26±0.18
SVHN	CRL-margin	96.65±0.07	4.41±0.20	3.85±0.18	42.91±0.99	**26.58±1.04**	**0.83±0.05**	**1.35±0.00**	**5.15±0.07**
DenseNet	MCdropout	96.82±0.04	5.10±0.52	4.59±0.51	39.57±2.58	31.04±1.67	0.42±0.06	1.29±0.03	4.97±0.11
	Aleatoric+MC	**96.86±0.14**	5.68±1.19	5.18±1.15	39.09±2.28	31.43±3.61	0.79±0.87	1.44±0.35	5.18±1.15
	AES(k=10)	96.78±0.08	4.50±0.16	3.98±0.15	**43.43±1.39**	26.16±1.17	0.41±0.09	1.24±0.02	4.96±0.10
	AES(k=30)	96.80±0.07	**4.29±0.14**	**3.77±0.13**	43.14±1.30	**25.86±0.84**	**0.34±0.07**	**1.21±0.02**	**4.90±0.10**



(a) CIFAR-10



(b) CIFAR-100



(c) SVHN

Figure S1. Risk-coverage curves from PreAct-ResNet110 on (a) CIFAR-10, (b) CIFAR-100, and (c) SVHN.

Table S4. Comparison of ensembles of five classifiers. λ is set to 0.5 for CRL models. For each experiment, the best result is shown in boldface. AURC and E-AURC values are multiplied by 10^3, and NLL are multiplied by 10 for clarity. All remaining values are percentage.

Dataset Model	Method	ACC (↑)	AURC (↓)	E-AURC (↓)	AUPR-Err (↑)	FPR-95% TPR (↓)	ECE (↓)	NLL (↓)	Brier (↓)
CIFAR-10 VGG-16	Baseline	95.02	4.45	3.19	**46.45**	**33.73**	1.52	1.92	7.65
	CRL-entropy	94.81	5.06	3.69	45.96	34.68	0.97	1.79	7.77
	CRL-softmax	**95.09**	**4.32**	**3.09**	45.27	37.88	1.32	1.78	**7.51**
	CRL-margin	94.85	5.05	3.70	42.01	40.77	**0.93**	**1.71**	7.67
CIFAR-10 ResNet110	Baseline	95.42	4.01	2.95	**44.14**	**29.03**	1.12	1.63	6.86
	CRL-entropy	95.15	4.12	2.93	43.38	34.02	**0.42**	1.50	7.22
	CRL-softmax	**95.55**	**3.72**	**2.72**	44.01	29.88	0.84	1.50	**6.60**
	CRL-margin	95.23	4.26	3.10	37.90	39.83	0.76	**1.46**	7.03
CIFAR-10 DenseNet	Baseline	**96.03**	**3.02**	**2.22**	44.17	30.73	0.79	1.29	**5.97**
	CRL-entropy	95.89	3.33	2.47	42.80	33.57	0.57	1.32	6.31
	CRL-softmax	95.97	3.17	2.35	45.25	29.77	0.85	**1.27**	5.99
	CRL-margin	95.50	3.45	2.43	**47.12**	**28.88**	**0.45**	1.32	6.48
CIFAR-100 VGG-16	Baseline	78.34	54.53	29.16	64.99	58.44	4.07	9.53	31.05
	CRL-entropy	78.43	55.19	30.05	64.50	60.36	3.85	9.14	30.86
	CRL-softmax	**78.53**	**52.53**	**27.63**	**66.53**	**57.89**	**3.80**	9.11	**30.47**
	CRL-margin	77.84	58.27	31.67	61.69	63.94	4.42	**9.08**	30.84
CIFAR-100 ResNet110	Baseline	78.83	54.91	30.72	64.42	58.99	2.39	8.63	30.19
	CRL-entropy	78.69	54.49	29.97	64.51	58.51	**1.95**	8.31	30.01
	CRL-softmax	**79.08**	**52.87**	**29.27**	**64.88**	**57.74**	2.11	**8.06**	**29.59**
	CRL-margin	79.01	57.20	33.44	56.87	68.41	2.04	**8.06**	29.90
CIFAR-100 DenseNet	Baseline	80.34	47.43	26.70	**63.83**	56.10	1.87	7.43	27.74
	CRL-entropy	80.47	46.10	**25.65**	63.73	**55.65**	1.81	7.20	27.47
	CRL-softmax	**80.85**	**45.63**	25.99	61.46	57.33	1.79	**7.13**	**27.34**
	CRL-margin	80.29	48.15	27.30	59.93	63.01	**1.53**	7.20	27.60
SVHN VGG-16	Baseline	96.91	4.48	4.00	**40.66**	28.64	1.09	1.60	4.93
	CRL-entropy	**97.01**	**3.96**	**3.51**	39.80	**27.02**	**0.78**	**1.30**	**4.75**
	CRL-softmax	96.95	4.07	3.60	40.52	29.25	1.02	1.53	4.92
	CRL-margin	96.84	4.30	3.80	37.62	30.04	0.86	1.42	4.92
SVHN ResNet110	Baseline	97.13	4.33	3.91	**42.52**	26.30	0.92	1.38	4.47
	CRL-entropy	97.24	**3.56**	**3.17**	41.58	**25.80**	**0.59**	**1.13**	4.30
	CRL-softmax	97.29	3.80	3.43	40.75	26.80	0.88	1.23	4.26
	CRL-margin	**97.31**	3.61	3.24	36.75	27.03	0.72	1.16	**4.24**
SVHN DenseNet	Baseline	**97.24**	4.93	4.55	36.49	30.54	0.83	1.34	4.51
	CRL-entropy	97.15	3.85	3.44	40.59	**27.16**	0.72	**1.17**	4.47
	CRL-softmax	97.18	4.10	3.70	**43.31**	29.05	0.87	1.25	4.46
	CRL-margin	97.19	**3.73**	**3.34**	35.40	27.98	**0.59**	1.18	**4.41**

Table S5. Comparison of ensembles of five classifiers. λ is set to 1 for CRL models. For each experiment, the best result is shown in boldface. AURC and E-AURC values are multiplied by 10^3, and NLL are multiplied by 10 for clarity. All remaining values are percentage.

Dataset Model	Method	ACC (↑)	AURC (↓)	E-AURC (↓)	AUPR-Err (↑)	FPR-95% TPR (↓)	ECE (↓)	NLL (↓)	Brier (↓)
CIFAR-10 VGG-16	Baseline	**95.02**	**4.45**	**3.19**	46.45	**33.73**	1.52	1.92	**7.65**
	CRL-entropy	94.70	5.12	3.69	43.88	37.92	**0.50**	1.86	7.77
	CRL-softmax	94.60	5.21	3.72	**46.80**	37.22	1.32	**1.71**	8.03
	CRL-margin	94.77	5.67	4.28	36.91	47.22	0.99	1.90	8.15
CIFAR-10 ResNet110	Baseline	**95.42**	**4.01**	**2.95**	44.14	**29.03**	1.12	1.63	**6.86**
	CRL-entropy	95.16	4.42	3.23	39.56	35.95	1.68	1.63	7.43
	CRL-softmax	94.70	4.58	3.15	**45.23**	34.15	0.72	**1.53**	7.71
	CRL-margin	94.62	4.91	3.44	41.74	35.50	**0.68**	1.58	7.87
CIFAR-10 DenseNet	Baseline	**96.03**	**3.02**	**2.22**	**44.17**	**30.73**	0.79	**1.29**	**5.97**
	CRL-entropy	95.52	3.72	2.70	43.82	32.14	1.50	1.45	6.73
	CRL-softmax	95.34	3.92	2.81	43.89	32.61	**0.52**	1.40	6.94
	CRL-margin	95.18	4.26	3.08	40.61	37.75	0.61	1.45	7.28
CIFAR-100 VGG-16	Baseline	78.34	54.53	29.16	64.99	58.44	4.07	9.53	31.05
	CRL-entropy	**78.66**	55.05	28.46	65.20	59.04	2.17	8.59	**29.96**
	CRL-softmax	78.09	**53.74**	**27.76**	**67.01**	**56.86**	2.76	**8.48**	30.29
	CRL-margin	78.08	58.63	32.63	62.32	62.04	**2.14**	8.67	30.52
CIFAR-100 ResNet110	Baseline	78.83	54.91	30.72	64.42	58.99	2.39	8.63	30.19
	CRL-entropy	78.56	53.92	29.09	64.32	58.53	2.39	8.63	30.19
	CRL-softmax	78.40	**53.55**	**28.33**	**66.35**	**56.43**	2.38	7.93	30.04
	CRL-margin	**78.84**	55.85	31.69	58.53	66.82	**1.78**	**7.61**	**29.69**
CIFAR-100 DenseNet	Baseline	80.34	47.43	26.70	**63.83**	**56.10**	1.87	7.43	27.74
	CRL-entropy	80.18	47.37	26.29	62.65	56.91	2.18	7.21	27.73
	CRL-softmax	80.38	**46.63**	**25.98**	62.59	58.81	**1.45**	6.95	27.43
	CRL-margin	**80.50**	48.27	27.88	57.82	63.64	1.55	**6.94**	**27.42**
SVHN VGG-16	Baseline	96.91	4.48	4.00	40.66	28.64	1.09	1.60	4.93
	CRL-entropy	**96.98**	4.16	3.70	**41.49**	**26.62**	**0.45**	**1.30**	**4.75**
	CRL-softmax	**96.98**	**4.02**	**3.56**	41.21	28.95	0.81	**1.30**	4.79
	CRL-margin	96.97	4.05	3.59	38.50	29.18	0.47	1.46	4.87
SVHN ResNet110	Baseline	97.13	4.33	3.91	**42.52**	26.30	0.92	1.38	4.47
	CRL-entropy	**97.31**	**3.51**	**3.15**	37.65	28.08	0.60	1.13	4.30
	CRL-softmax	97.26	3.82	3.44	40.00	26.58	0.56	**1.12**	4.33
	CRL-margin	97.26	3.66	3.28	37.61	**25.17**	**0.50**	1.14	**4.27**
SVHN DenseNet	Baseline	**97.24**	4.93	4.55	36.49	30.54	0.83	1.34	4.51
	CRL-entropy	97.18	**3.70**	**3.30**	39.74	26.43	0.74	1.16	**4.44**
	CRL-softmax	97.13	3.85	3.44	39.91	**25.77**	**0.53**	**1.14**	4.46
	CRL-margin	97.19	3.76	3.37	**40.02**	28.49	0.81	1.17	4.53

Table S6. Performances of CRL models on out-of-distribution detection task. The means and standard deviations are computed from five models trained to evaluate the ordinal ranking performance. For each comparison, better result is shown in boldface. All values are percentage.

In-dist Model	Out-of-dist	FPR-95%TPR(↓)	Detection Err.(↓)	AUROC(↑)	AUPR-In(↑)	AUPR-Out(↑)
				Baseline / CRL **Baseline+ODIN / CRL+ODIN** **Baseline+Mahalanobis / CRL+Mahalanobis**		
SVHN ResNet110	TinyImageNet	29.65±2.40 / **5.89±0.70**	12.11±0.96 / **5.05±0.23**	93.00±1.06 / **98.83±0.13**	96.31±0.91 / **99.56±0.04**	84.95±1.41 / **96.72±0.50**
		27.50±3.09 / **2.17±0.26**	13.14±1.29 / **3.41±0.23**	92.32±1.38 / **99.39±0.08**	95.63±1.17 / **99.76±0.03**	85.61±1.85 / **98.41±0.30**
		0.24±0.08 / 0.30±0.12	**1.15±0.16** / 1.39±0.27	**99.88±0.03** / 99.82±0.06	**99.96±0.01** / 99.93±0.02	**99.39±0.19** / 98.99±0.28
	LSUN	32.37±2.78 / **7.48±0.91**	13.01±1.17 / **5.50±0.20**	92.19±1.39 / **98.62±0.17**	95.82±1.30 / **99.49±0.06**	83.48±1.74 / **96.14±0.62**
		29.57±3.98 / **2.92±0.51**	14.06±1.23 / **3.88±0.30**	91.56±1.78 / **99.28±0.11**	95.19±1.65 / **99.72±0.03**	84.42±2.38 / **98.06±0.42**
		0.08±0.05 / **0.06±0.07**	0.88±0.14 / **0.85±0.32**	**99.91±0.03** / 99.89±0.06	**99.97±0.01** / 99.96±0.02	**99.45±0.25** / 99.03±0.38
	iSUN	31.43±2.63 / **6.40±0.78**	12.67±1.15 / **5.17±0.23**	92.46±1.39 / **98.75±0.14**	96.36±1.21 / **99.58±0.04**	82.54±1.66 / **96.13±0.59**
		29.73±3.71 / **2.55±0.41**	13.87±1.43 / **3.67±0.26**	91.59±1.91 / **99.30±0.07**	95.67±1.58 / **99.76±0.02**	83.04±2.42 / **97.92±0.29**
		0.04±0.05 / **0.03±0.03**	**0.73±0.22** / 0.81±0.24	**99.89±0.03** / 99.88±0.04	99.97±0.04 / **99.97±0.01**	**99.31±0.26** / 98.79±0.63
SVHN DenseNet	TinyImageNet	26.32±5.55 / **7.99±2.49**	11.49±1.61 / **5.75±0.71**	93.75±1.43 / **98.53±0.41**	96.62±0.94 / **99.43±0.16**	87.30±2.74 / **96.15±1.16**
		19.93±4.43 / **3.39±1.34**	11.46±1.80 / **4.04±0.80**	94.06±1.47 / **99.17±0.28**	96.56±0.94 / **99.65±0.12**	90.03±2.39 / **98.00±0.67**
		1.44±1.62 / **1.03±1.41**	2.42±1.15 / **1.86±0.92**	99.37±0.91 / **99.48±0.79**	99.52±1.08 / **99.62±0.99**	98.45±1.04 / **98.48±0.80**
	LSUN	28.95±5.80 / **11.05±3.09**	12.39±1.84 / **6.58±0.74**	92.95±1.76 / **98.12±0.49**	96.11±1.23 / **99.29±0.19**	85.93±3.12 / **95.06±1.40**
		22.22±4.86 / **4.63±1.84**	12.35±1.98 / **4.68±0.94**	93.32±1.80 / **98.93±0.36**	96.15±1.22 / **99.56±0.15**	88.83±2.73 / **97.45±0.89**
		0.41±0.84 / 0.44±0.46	**1.23±0.63** / 1.23±0.66	99.73±0.60 / **99.75±0.17**	**99.86±0.88** / 99.79±0.13	**98.97±0.60** / 98.70±0.62
	iSUN	28.22±5.59 / **9.94±2.78**	12.08±1.62 / **6.43±0.73**	93.11±1.57 / **98.25±0.47**	96.57±0.96 / **99.39±0.16**	85.04±3.20 / **95.04±1.50**
		22.37±4.59 / **4.39±1.90**	12.20±1.82 / **4.51±1.03**	93.37±1.62 / **98.96±0.36**	96.51±0.97 / **99.61±0.14**	87.95±2.78 / **97.38±0.99**
		0.06±0.46 / 0.08±0.16	**0.81±0.62** / 1.13±0.26	**99.87±0.21** / 99.78±0.05	99.94±0.17 / **99.94±0.03**	**98.92±0.33** / 98.56±0.30
CIFAR-10 ResNet110	TinyImageNet	66.09±2.86 / **53.17±5.60**	22.59±1.81 / **22.06±2.35**	82.59±2.91 / **86.25±2.76**	79.63±5.39 / **86.56±3.27**	82.07±2.00 / **85.61±2.50**
		49.33±4.19 / **43.08±5.15**	22.08±2.28 / **17.69±2.02**	84.31±3.22 / **90.40±1.91**	80.73±5.07 / **90.77±2.13**	86.01±2.30 / **90.03±1.77**
		8.46±2.12 / 9.44±2.25	**6.39±0.87** / 7.02±0.91	**98.34±0.41** / 97.92±0.41	**98.40±0.37** / 97.85±0.35	**98.22±0.49** / 98.02±0.44
	LSUN	57.65±2.89 / **44.53±6.57**	**17.78±1.21** / 17.89±1.87	88.25±1.54 / **90.46±1.93**	87.73±2.59 / **91.37±1.95**	87.06±1.29 / **89.60±1.97**
		34.72±5.75 / **32.10±5.29**	16.29±1.76 / **13.50±1.64**	90.63±1.97 / **93.90±1.23**	88.98±2.79 / **94.48±1.21**	91.47±1.66 / **93.30±1.25**
		6.33±2.54 / **5.52±1.39**	5.51±1.25 / **5.16±0.76**	98.66±0.51 / **98.71±0.29**	**98.79±0.48** / 98.76±0.25	98.49±0.61 / **98.62±0.31**
	iSUN	61.78±2.67 / **50.34±5.87**	20.07±1.47 / **20.01±2.12**	85.84±2.15 / **88.39±2.27**	85.78±3.71 / **90.25±2.28**	83.44±1.63 / **86.29±2.37**
		41.95±4.35 / **38.08±5.21**	19.02±1.70 / **14.89±1.58**	87.95±2.39 / **92.66±1.31**	86.90±3.59 / **93.96±1.17**	88.08±1.90 / **91.01±1.48**
		8.19±2.47 / 9.48±1.66	**6.53±1.08** / 7.07±0.71	**98.10±0.57** / 97.97±0.37	**98.29±0.50** / 98.10±0.33	**97.89±0.72** / 97.81±0.42
CIFAR-10 DenseNet	TinyImageNet	45.81±3.95 / **29.87±4.09**	13.15±1.41 / **12.99±1.03**	93.25±1.04 / **94.50±0.84**	94.53±0.94 / **95.17±0.71**	91.82±1.24 / **93.87±1.03**
		10.73±6.24 / **10.41±3.09**	7.09±2.02 / **6.89±1.10**	97.86±1.09 / **97.97±0.56**	97.90±1.04 / **98.16±0.48**	**97.84±1.13** / 97.78±0.65
		6.99±1.13 / **6.28±3.18**	5.92±0.58 / **5.61±1.54**	98.37±0.50 / **98.52±1.17**	**98.22±1.29** / 98.09±2.07	98.49±0.38 / **98.57±0.82**
	LSUN	36.31±3.64 / **21.22±2.73**	10.60±0.88 / **10.59±0.55**	95.18±0.64 / **96.34±0.44**	96.16±0.50 / **96.80±0.35**	94.14±0.86 / **95.94±0.57**
		4.32±2.55 / 5.29±1.53	**4.46±1.15** / 5.03±0.71	**99.04±0.46** / 98.81±0.29	**99.10±0.41** / 98.92±0.24	**98.99±0.50** / 98.70±0.35
		5.27±1.15 / **3.86±2.15**	5.08±0.57 / **4.26±1.11**	98.73±0.50 / **98.89±0.66**	**98.68±1.56** / 98.67±1.07	98.71±0.36 / **98.91±0.50**
	iSUN	39.84±4.40 / **25.59±3.52**	**11.49±1.26** / 11.75±0.84	94.51±0.91 / **95.52±0.67**	95.99±0.70 / **96.44±0.50**	92.61±1.26 / **94.52±0.88**
		6.61±4.00 / 7.54±2.08	**5.49±1.41** / 6.03±0.92	**98.67±0.70** / 98.45±0.41	**98.87±0.55** / 98.73±0.33	**98.46±0.87** / 98.13±0.54
		6.49±1.07 / **6.20±1.45**	5.67±0.52 / **5.54±0.69**	**98.52±0.21** / 98.49±0.51	98.34±0.30 / **98.38±0.93**	98.44±0.19 / **98.50±0.41**

Table S7. Comparison of five sampling strategies for ResNet18 on CIFAR datasets. The means and standard deviations of accuracy values over five runs are reported. For each experiment, the best result is shown in boldface. The percentage in parentheses next to the stage number indicates the proportion of the labeled dataset to the entire training dataset (i.e., 50,000).

Dataset	Sampling	Stage									
		1st (4%)	**2nd** (8%)	**3rd** (12%)	**4th** (16%)	**5th** (20%)	**6th** (24%)	**7th** (28%)	**8th** (32%)	**9th** (36%)	**10th** (40%)
CIFAR-10	random	64.86±1.43	77.35±0.69	82.31±1.42	85.43±0.96	86.84±0.71	88.36±0.82	89.45±0.70	90.47±0.54	91.10±0.57	91.50±0.58
	entropy	64.98±1.12	80.40±0.56	85.69±0.59	88.30±0.34	90.14±0.28	91.62±0.19	92.69±0.13	93.35±0.15	93.82±0.26	94.32±0.18
	coreset	65.36±0.95	79.55±0.39	84.79±0.23	87.58±0.24	89.42±0.19	91.00±0.24	91.94±0.34	92.70±0.21	93.48±0.16	93.81±0.20
	MC-entropy	59.22±1.89	75.54±1.46	84.62±0.59	87.93±0.52	90.10±0.41	91.55±0.22	92.72±0.17	93.32±0.19	**93.96±0.17**	**94.33±0.17**
	CRL-softmax	**65.91±1.44**	**80.60±0.58**	**85.84±0.36**	**89.00±0.06**	**90.82±0.16**	**91.80±0.29**	**92.78±0.10**	**93.38±0.24**	93.82±0.14	94.09±0.09
CIFAR-100	random	20.57±0.49	31.72±1.36	42.61±1.45	48.84±0.35	54.30±0.41	58.34±0.50	61.11±0.16	63.39±0.28	65.19±0.28	66.73±0.28
	entropy	19.86±0.42	32.41±0.17	42.63±1.30	50.45±0.80	55.36±0.47	60.10±0.45	63.22±0.50	65.54±0.46	68.01±0.42	69.35±0.23
	coreset	20.27±0.64	**33.47±0.74**	**44.36±1.20**	50.31±0.46	56.00±0.65	59.34±0.46	62.53±0.43	64.82±0.27	66.66±0.37	68.20±0.28
	MC-entropy	19.45±0.70	31.20±2.18	41.57±1.18	50.03±0.69	56.18±0.60	60.28±0.34	63.55±0.20	66.31±0.22	68.25±0.35	69.97±0.40
	CRL-softmax	**20.72±0.34**	32.17±1.03	43.87±1.82	**51.22±0.99**	**57.69±0.79**	**61.65±0.31**	**64.27±0.36**	**66.71±0.57**	**68.78±0.27**	**70.40±0.32**

References

Beluch, W. H., Genewein, T., Nürnberger, A., and Köhler, J. M. The power of ensembles for active learning in image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Corbière, C., THOME, N., Bar-Hen, A., Cord, M., and Pérez, P. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*. 2019.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*. 2017.

Geifman, Y., Uziel, G., and El-Yaniv, R. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*, 2019.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. Snapshot ensembles: Train 1, get M for free. In *International Conference on Learning Representations*, 2017.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*. 2017.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*. 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Naeini, M. P., Cooper, G. F., and Hauskrecht, M. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Yoo, D. and Kweon, I. S. Learning loss for active learning. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Yu, F., Zhang, Y., Song, S., Seff, A., and Xiao, J. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365, 2015.

Zhang, Z., Dalca, A. V., and Sabuncu, M. R. Confidence calibration for convolutional neural networks using structured dropout. *CoRR*, abs/1906.09551, 2019.