Flexible Visual Recognition by Evidential Modeling of Confusion and Ignorance

Lei Fan[1], Bo Liu[2], Haoxiang Li[2], Ying Wu[1] and Gang Hua[2]
[1]Northwestern University [2]Wormpex AI Research
leifan@u.northwestern.edu,yingwu@northwestern.edu,{richardboliu,lhxustcer,ganghua}@gmail.com

Abstract

In real-world scenarios, typical visual recognition systems could fail under two major causes, i.e., the misclassification between known classes and the excusable misbehavior on unknown-class images. To tackle these deficiencies, flexible visual recognition should dynamically predict multiple classes when they are unconfident between choices and reject making predictions when the input is entirely out of the training distribution. Two challenges emerge along with this novel task. First, prediction uncertainty should be separately quantified as confusion depicting inter-class uncertainties and ignorance identifying out-of-distribution samples. Second, both confusion and ignorance should be comparable between samples to enable effective decision-making. In this paper, we propose to model these two sources of uncertainty explicitly with the theory of Subjective Logic. Regarding recognition as an evidence-collecting process, confusion is then defined as conflicting evidence, while ignorance is the absence of evidence. By predicting Dirichlet concentration parameters for singletons, comprehensive subjective opinions, including confusion and ignorance, could be achieved via further evidence combinations. Through a series of experiments on synthetic data analysis, visual recognition, and open-set detection, we demonstrate the effectiveness of our methods in quantifying two sources of uncertainties and dealing with flexible recognition.

1. Introduction

When employing visual classifiers in open-world conditions, obtaining reliable uncertainty estimations could significantly benefit downstream tasks, including autonomous driving [1, 13, 37], medical diagnosis [36, 49], and embodied intelligence [40, 35]. Recent uncertainty quantification techniques [53, 15, 28, 32, 34, 33, 9, 43] have achieved notable progress toward this goal by saying *"I do not know"* when the testing distributions differ from the training. However, besides directly giving no prediction, a more flexible and informative visual recognition system could also give combined predictions when possible, implying the correct answer is one of its predictions but uncertain. Naturally, the capability of rejecting or providing unspecific predictions demands separately measuring different sources of uncertainties, *i.e.*, ignorance and confusion, if seen from the Subjective Logic [21] perspective. Furthermore, to enable flexible recognition, both uncertainties should possess the virtue of comparability between samples and in-sample additivity.

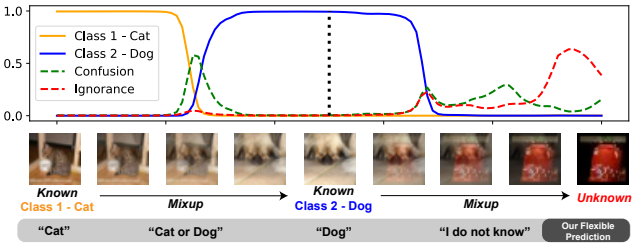


Figure 1: Classification of the proposed approach on images interpolated from a *known-known-unknown* triplet. Ignorance reflects the lack of evidence, whereas confusion is caused by conflicting evidence, *i.e.*, evidence that fails to provide discrimination between specific classes. A flexible visual recognition system could provide combined predictions when having large confusion and reject making predictions for unknown-class samples. Note the mixup images are for illustrative purposes and are not a requisite in our training.

In evidential deep learning, the training of a recognition model could be regarded as an evidence-collecting process [21, 44]. Unlike ignorance describing a total lack of evidence, confusion is defined as conflicting evidence, which mandates the existence of multiple hypotheses in the frame of discernment. In other words, we cannot assess confusion for a single-class classification problem. The mass of confusion between two classes then reflects shared features that contribute to both classes while not discriminative. Likewise, confusion exists for all combinations of classes larger than two. Unlike typical visual classifiers that only predict

a single output, flexible predictions could be obtained if we could combine singleton belief derived from class-exclusive evidence with their inter-class confusion.

With great potential for explicitly estimating confusion and ignorance, this area is still under-explored for deep visual classifiers. Recent methods regard uncertainties as the degree of mismatch between training and testing distributions, which comprise but do not distinguish between confusion and ignorance. Deep Bayesian models, including dropout [15, 22] and ensemble-based approximations [28, 51, 4], require multiple forwards to estimate the posterior predictive distribution. Evidential models [44, 1, 11, 5] predict parameters of the posterior of class distribution directly. However, these models regard uncertainty as a whole term covering both confusion and ignorance, making it infeasible to perform flexible visual recognition further.

Distinct from existing uncertainty quantification methods, the proposed method models confusion and ignorance for each sample separately, which provides valuable information to facilitate various visual tasks, including flexible visual recognition. An illustrative example with the prediction of our method is shown in Fig. 1. Under the theory of Subjective Logic [6, 21], confusion is defined as the shared evidence contributing to multiple categories while not discriminative between them, while ignorance is completely missing evidence.

The contribution of this paper could be summarized as follow: (1) The proposed method could explicitly predict two sources of uncertainties, *i.e.*, confusion and ignorance, simultaneously for each sample. (2) The solution to confusion and ignorance is based on standard architectures, and the training does not rely on external information. (3) The effectiveness of the proposed method is extensively validated across different experiments, including studies on synthetic data, visual recognitions, and open-set detections.

2. Related Work

Uncertainty estimation. Typical neural networks can not detect their own failure. However, this ability can be important in several real-world applications, like rejecting unseen samples, and providing prediction confidence, to name a few. Bayesian NN [15, 27, 47, 14, 25] predicts epistemic uncertainty as the mutual information between model parameters and samples. By assuming a probabilistic prior on the network, it approximates prediction variance by sampling weight during inference. Several works [20, 49, 23] choose to model epistemic and aleatoric uncertainties separately. The Subjective Logic on which our method is established falls within the realm of epistemology instead of a frequentist (aleatoric) view. In other words, we focus on further separating epistemic uncertainty into confusion and ignorance.

Evidential deep learning [44, 1, 5, 11], in contrast, proposes to learn the prior of the predictions directly. The prior, known as evidential prior, is interpreted as beliefs in Dempster-Shafer Threory [6]. In [11], they model the first- and second-order uncertainties by introducing an auxiliary uncertainty network to approximate the difference between Dirichlet distributions.

While uncertainty is provided to describe the variance of model prediction, it is not clear if uncertainty comes from different sources when dealing with in-distribution or out-of-distribution data. Orthogonal to previous approaches, we separate the uncertainty into confusion and ignorance in this work. Confusion depicts the uncertainty between different known classes, while ignorance decides whether the sample is unknown. With this separation, we can make dynamic predictions on known classes and reject unknown classes at the same time.

Open Set Recognition. Machine learning models are usually designed with the closed-set assumption, where testing data shares the same distribution as the training. Open-set recognition (OSR) [8] introduces semantic shifts to the problem. Samples are from the classes that are not in the training set. Out-of-distribution (OOD) [19] detection introduces domain shifts to the testing set. In both of the settings, models should have the ability to reject unknown samples.

In general, OSR and OOD methods reject unknown samples depending on reweighting outputs [8, 16, 24, 19, 30, 50], getting better feature embedding metrics [31, 10, 42, 38], and exploring reconstruction errors [39, 46, 45, 52, 41]. These metrics are all related to the quality of classifier prediction. However, the recognition can fail on closed-set samples because of the existence of confusion. In this work, we show that when the confusion between known classes is adequately modeled, unknown samples can be identified more accurately.

Conformal Prediction. Parallel to our task, conformal prediction is a paradigm that could provide single or multiple predictions by empirically constructing confidence regions [2, 3, 48]. However, conformal prediction is confined to closed scenarios without open samples, as the empirical quantile is established on a labeled validation set sampled from the same testing distribution.

3. Flexible Recognition

Flexible recognition aims to provide a classification model $\mathcal{M}$ that could deliver adaptive predictive sets. Specifically, the model rejects samples, *i.e.*, making no prediction, when the input is entirely out of the training distribution, like an open-set sample. The model also should cautiously give a set of predictions with the true class contained in it when being unsure.

For a K-class classification problem, we formalize the flexible recognition system $\mathcal{M}(\cdot)$ as $\{y_1, \ldots, y_k\} = \mathcal{M}(\mathbf{x})$ where $\mathbf{x}$ denotes the input image, and the predictive set

$\{y_1, \ldots, y_k\}$ obeys $0 \leq k \leq K$. Therefore, $k = 0$ means the recognition system rejects making a prediction, and the true label $\mathbf{y}$ is supposed to be contained in the predictive set when $\mathbf{x}$ is from known classes.

4. Method

In this paper, we propose to tackle flexible recognition by separately estimating the confusion and ignorance for each sample. Intuitively, confusion denotes conflicting evidence between known classes, which should be additive to single-class beliefs for making reasonable multiple predictions. Ignorance denotes a lack of evidence to support rejecting samples. Most of this section falls into the proposed evidential modeling, which formulates confusion and ignorance in visual recognition under the theory of Subject Logic [21]. The approach to combining evidence and developing opinions follows. The approach to achieve flexible recognition is presented in the final.

4.1. Preliminaries

Existing learning-based visual recognition models often rely on a softmax layer to give class probabilities. As a point estimation of predictive distribution, the classifier trained with the cross-entropy loss tends to deliver inflating probabilities to a single class and could not provide a reliable estimation of uncertainties [19].

We develop confusion and ignorance based on the Dempster-Shafer Theory of Evidence (DST) [6], which is a generalized scheme towards subjective probabilities [21]. The theory allows plausible reasoning with operations of evidence, namely the combination of evidence. Consider a K-class recognition task, $\Theta = \{i, 1 \leq i \leq K\}$ would be the frame of discernment containing exclusive propositions, *e.g.*, class labels. Extensively, the general propositions under this frame would be the set of all subsets of Θ, which is

$$2^{\Theta} = \{\emptyset, 1, \ldots, K, \{1, 2\}, \ldots, \Theta\}, \tag{1}$$

where 2^{Θ} contains a total of 2^K elements.

Supposing $b_A \in [0, 1]$ as a measure of belief mass contained in proposition A, the total mass of general propositions satisfies $\sum_{A \in 2^{\Theta}} b_A = 1$. The belief for any proposition is then defined as the summation of contained mass, which is formulated as $b_A = \sum_{B \subseteq A} b_B$.

And it is worth noting that $B \subseteq A$ suggests the logical statement that B implies A. The total belief in A is the sum of belief in all propositions that imply A plus the belief in A itself. For a more intuitive understanding, considering a binary visual classification task, the belief for predicting both classes is the combination of mass shared between classes and also class-exclusive masses.

We could further define the plausibility of A as pl_A, which is the total mass of propositions that has a non-empty union with the current one. Also, the plausibility of a hypothesis could only be larger or equal to its own belief, *i.e.*, $pl_A \geq b_A$. We demonstrate the basic probability assignments and their relations in Fig. 2.

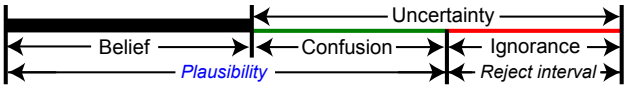


Figure 2: The relation between different masses constituting the final set of opinions towards a hypothesis.

4.2. Uncertainty, Confusion, and Ignorance

Recent evidential deep learning methods [44] develop uncertainties under multinomial opinions, which model belief for singletons as $\{b_i, i = 1, \ldots, K\}$ and regard the leftover as the total uncertainty $\mathcal{U}$. We, therefore, have $\mathcal{U} + \sum_{i=1}^{K} b_i = 1$.

However, in our modeling, we argue the uncertainty $\mathcal{U}$ for each sample $\mathbf{x}$ comes from two distinct sources, *i.e.*, confusion $\mathcal{C}$ and ignorance $\mathcal{I}$, which is written as

$$\mathcal{U}^{\mathbf{x}} = \mathcal{C}^{\mathbf{x}} + \mathcal{I}^{\mathbf{x}}. \tag{2}$$

Intuitively, a large confusion $\mathcal{C}^{\mathbf{x}}$ denotes the model is hard to distinguish $\mathbf{x}$ from known classes. For example, the model could have high confusion with a huskie image when the model has known classes of *dog* and *wolf*. An image from previously unknown classes, on the other hand, could have high ignorance. The superscript $\mathbf{x}$ will be omitted in the following for clarity.

To introduce the separate measurements of confusion and ignorance, our method is formalized with hyper opinions, composing masses of 2^K subsets for a K-class frame of discernment. Therefore, the overall confusion $\mathcal{C}$ is the total mass of the non-singleton subsets as $\mathcal{C} = \sum_{A, A \in 2^{\Theta}, 2 \leq |A| \leq K} b_A$. In other words, the confusion $\mathcal{C}$ is the sum of masses shared between two or more classes. Altogether, following Eq. 2, we have 2^K mass values satisfy

$$\mathcal{C} + \mathcal{I} + \sum_{i=1}^{K} b_i = \sum_{A, A \in 2^{\Theta}, 2 \leq |A| \leq K} b_A + \mathcal{I} + \sum_{i=1}^{K} b_i = 1, \tag{3}$$

where $\mathcal{I} \geq 0$ and $b \geq 0$ for all singletons and non-singleton subsets. The ignorance $\mathcal{I}$, therefore, could be regarded as the mass placed on the empty set $\emptyset$ in the frame, which indicates the level of lacking evidence. Confusion, defined on subsets with cardinalities larger than 1, reflects evidence that fails to discriminate between specific singletons.

To further facilitate our evidence combination process, we group non-singleton confusion terms based on whether they hold evidence for a particular class i.

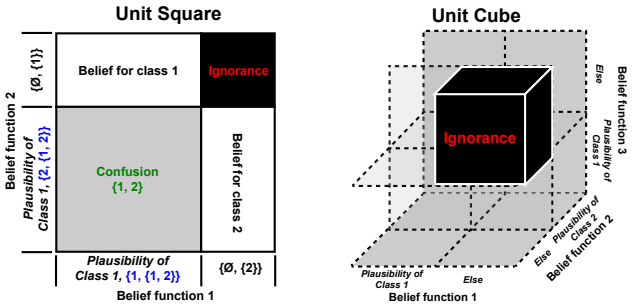


Figure 3: Two graphical demonstrations of evidence combination on 2- and 3-class classification task. We only show ignorance in the 3-class example to avoid overlappings.

Class i-related confusion mass $\mathcal{C}_i$. It is defined as $\mathcal{C}_i = \sum_{A,A\in 2^\Theta,|A|\geq 2,i\cap A=i} b_A$. To be more specific, $\mathcal{C}_i$ is the total mass placed on the set of all confusion terms that are supersets of singleton i.

Class i-unrelated confusion mass $\mathcal{C}_{\neg i}$. Conversely, we have $\mathcal{C}_{\neg i} = \sum_{A,A\in 2^\Theta,|A|\geq 2,i\cap A\neq i} b_A$. Accordingly, for any class i, we have

$$\mathcal{C} = \mathcal{C}_i + \mathcal{C}_{\neg i}. \tag{4}$$

And, as shown in Fig. 2, the plausibility of the current sample belonging to class i is $pl_i = b_i + \mathcal{C}_i$, which stands for a combination of class-exclusive singleton belief mass and class-shared confusion masses.

4.3. Evidence Combinations

The objective of the proposed approach is to predict ignorance and confusion explicitly for each sample. However, unlike ignorance which is a single term, the number of confusion reaches the exponential of classes, which means our model is required to give a comprehensive quantification of 2^K estimates. While intimidating, we propose to handle this by decomposing the problem into K plausibility functions $f_i(\cdot)$ for $i = 1, \ldots, K$ on the same frame. Each plausibility function $f_i(\cdot)$ is designed to give two predictions only considering class i, which is written as $f_i(\mathbf{x}) = (pl_i, 1 - pl_i)$.

Supposing we have obtained the plausibilities for all classes, any propositions, including the singleton class belief, confusion, and ignorance could then be derived by the rule of evidence combinations. In general, for K plausibility functions, the belief assignment of any proposition A is combined by computing as

$$b_A = \sum_{B,\cap B=A} \prod_{1\leq j\leq K} b_{B,j}(\mathbf{x}) = \sum_{B,\cap B=A} \prod_{1\leq j\leq K} f_j^B(\mathbf{x}). \tag{5}$$

Note we do not have the normalization term in our formulation because we do not exclude the empty set from our frame. We demonstrate the combination process with 2- and 3-class classification examples in Fig. 3. To clarify further, the singleton belief for class i is computed as

$$b_i = pl_i \prod_{1\leq j\leq K, j\neq i} (1 - pl_j) = f_i^1(\mathbf{x}) \prod_{1\leq j\leq K, j\neq i} f_j^2(\mathbf{x}). \tag{6}$$

And we have the total uncertainty $\mathcal{U} = 1 - \sum_{i=1}^K b_i$. As the ignorance $\mathcal{I}$ could be calculated similarly as

$$\mathcal{I} = \prod_{1\leq j\leq K} (1 - pl_j) = \prod_{1\leq j\leq K} f_j^2(\mathbf{x}), \tag{7}$$

the total confusion between all different class combinations is $\mathcal{C} = \mathcal{U} - \mathcal{I}$. The confusion term between specific classes could be further calculated with Eq. 5.

Intuitively, the belief for each combination could be regarded as the occupation in a unit K-dim volume. The modeling of confusion becomes feasible by spanning a K-dim hypothesis space with K plausibility functions. Moreover, the computation complexity of each combination is $\mathcal{O}(n)$, and for specific conditions, we could only calculate necessary confusion terms.

4.4. Developing Opinions

In this section, we describe how to develop opinions from training data. Each plausibility function $f_i(\cdot)$ can be constructed as a normalized dual-output linear layer or a single multi-output layer after being activated by a sigmoid function $\sigma(\cdot)$. In this work, the second is implemented to reduce the network parameters. In particular, the output is regarded as the value of class plausibility. The plausibility function is then formulated as

$$(pl_i, 1 - pl_i) = f_i(\mathbf{x}) = \sigma(w_i^\top \Phi(\mathbf{x})) \tag{8}$$

where $\Phi : \mathcal{X} \to \mathbb{R}^D$ is a feature embedding function.

Typically, only one deterministic label is given for each image in a standard visual recognition dataset. Following EDL [44], the learning of singleton belief is implemented as evidence acquisition on a Dirichlet prior. The loss of EDL is

$$\mathcal{L}_{\text{EDL}} = \sum_{i=1}^K y_i [\log(\sum_{j=1}^K \alpha_j) - \log(\alpha_i)], \tag{9}$$

where $\mathbf{y} = [y_1, y_2, \ldots, y_i, \ldots, y_K]^T$ is one-hot class label for a sample $\mathbf{x}$, and $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots \alpha_K]^T$ are parameters of a Dirichlet distribution $Dir(\cdot|\boldsymbol{\alpha})$.

Different from EDL, class evidence is replaced with belief. Hence, $\boldsymbol{\alpha}$ is directly calculated from singleton beliefs and overall uncertainty. It is derived as

$$\alpha_i = \frac{K b_i}{\mathcal{U}} + 1 = \frac{K b_i}{1 - \sum_{j=1}^K b_j} + 1, \tag{10}$$

where b_i could be obtained from Eq. 6. During inference, all opinions could be directly predicted by performing combinations on the output of plausibility functions.

To encourage the plausibility function to match our expected behavior, *i.e.*, predicting the plausibility instead of the belief of singleton, we add a regularization term as

$$\mathcal{L}_{\text{reg}} = \sum_{i=1}^{K} y_i [pl_i - (1 - \hat{\mathcal{I}})]^2, \tag{11}$$

where $\hat{\mathcal{I}}$ is the current estimation of ignorance.

Following EDL, a Kullback-Leibler loss is used to minimize evidence on unrelated classes as

$$\mathcal{L}_{\text{KL}} = KL(Dir(\cdot|\tilde{\boldsymbol{\alpha}})||Dir(\cdot|\langle 1, \ldots, 1\rangle)), \tag{12}$$

where $\tilde{\boldsymbol{\alpha}} = \mathbf{y} + (1 - \mathbf{y}) \odot \boldsymbol{\alpha}$, $\odot$ for element-wise multiplication. Combining all terms together yields the final loss as

$$\mathcal{L} = \mathcal{L}_{\text{EDL}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}}. \tag{13}$$

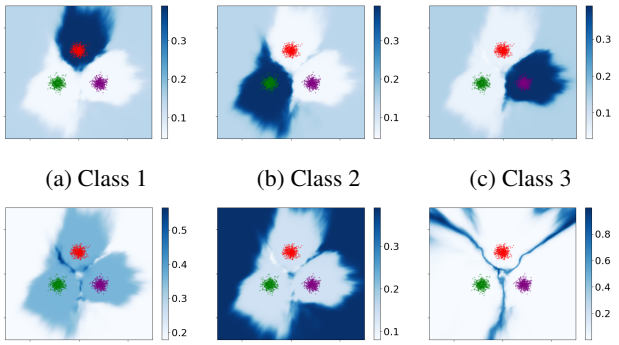
Each loss term is accompanied by a balance weight, and we gradually increase the effect of $\mathcal{L}_{kl}$ through an additional annealing coefficient.

After developing opinions with the proposed method, a straightforward solution would be setting the belief threshold for outputs to achieve flexible recognition. The sample will be rejected if the ignorance is too large that no combination would exceed the threshold. And the model gives incrementally combined predictions if no singleton belief meets the bar.

5. Experiments

The proposed method could model two sources of uncertainty for each sample to handle the task of flexible recognition. To enable better comparison with existing methods, our experiments are primarily decomposed into three components. (1) Demonstrating the separation of two sources of uncertainty, *i.e.*, confusion and ignorance. (2) Indicating the correct class on misclassified samples with estimated confusion. (3) Applying ignorance to compare with other methods on the task open-set detection. More experiments, including on adversarial-attacked samples and ablation studies, follow to give a more comprehensive evaluation.

Implementation details. We adopt the ResNet-18 as the backbone for our experiments except on synthetic data and open-set detection. The dimension of extracted feature is set to 512. For the proposed method, we apply the sigmoid activation on the last linear layer to work as our multiple plausibility functions. We empirically find both EDL [44] and our method are more sensitive to the learning rate. Specifically, we set the learning rate for both methods to 0.004 with a momentum of 0.9 for the batch size of 128. λ_{KL} in Eq. 13 anneals to 0 with epochs with the maximum coefficient of 0.05, and λ_{reg} is set to 1.



(a) Class 1 (b) Class 2 (c) Class 3

(d) Confusion (e) Ignorance (f) Standard Entropy

Figure 4: A 3-class classification problem. The Gaussian-distributed training data are depicted with dots, while the background color indicates the estimated value of the corresponding location. Our results are plotted from (a) to (e). The entropy over predictions of a standard net trained with cross-entropy loss is shown in (f) for comparison.

5.1. Synthetic Experiments

In this section, we examine the behavior of our approach in Fig. 4 in the 2-dimensional space. A dataset with three isotropic Gaussian distributed classes is created for training. The distances between each Gaussian are equal to 9, and we set $\sigma = 4$ for all Gaussian. Each class has 500 training samples which are denoted by colored dots. We discretize the background into 2D locations for testing. The background color denotes corresponding estimated values.

Both our method and the standard entropy are trained with the same multi-layer perceptron. As we can observe in Fig. 4 (a) to (e), the proposed method acts as a density estimator. Confusion happens between class boundaries, while ignorance is high for out-of-distribution data points. The proposed method does not require intricate adaptions to network architectures to achieve the desired properties. In Fig. 4 (f), the entropy of a standard deterministic neural net with the cross-entropy loss is plotted, which shows that using entropy to distinguish out-of-distributions could face multiple downsides due to its sharp boundaries.

5.2. Confusion on Misclassified Samples

In this part, the confusion is tested on misclassified samples, checking whether it is correlated with the ground-truth label. That is, the confusion should be high between the misclassified class and the ground-truth class. Besides in-sample comparison on the level of confusion, we argue the between-sample comparability of confusion is also critical for flexible visual recognition. For example, a flexible recognition system would tend to give a second prediction if its confusion is higher than in other samples. To address this concern, we employ the Area Under the Receiver Oper-

Dataset	CIFAR-10		CIFAR-100		Imagenet	
	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC
Softmax	95.2	63.4	76.0	57.3	54.3	58.2
EDL [44]	94.8	60.4	74.5	54.9	54.2	58.3
Dropout [15]	94.7	64.5	74.5	58.9	47.3	61.8
OvR	95.1	63.2	72.2	63.4	46.3	60.2
ASL [7]	95.2	70.9	75.8	79.9	54.1	62.8
Ours	95.0	**89.5**	74.9	**90.0**	54.6	**97.6**

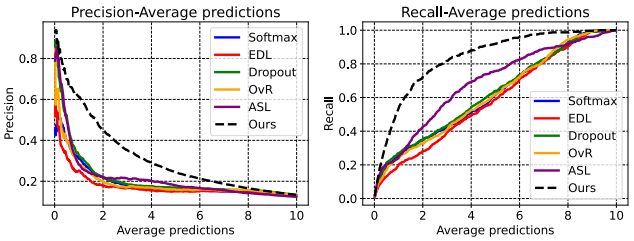
Table 1: Comparison on whether the confusion indicates the correct class on misclassified samples. Results of classification accuracy and AUROC on misclassified samples are shown on three datasets with different class scales.

ating Characteristic curve (AUROC) as our evaluation metric, which sorts the predicted value along samples in each class. More specifically, the ROC curve is a graph showing the true positive rate against the false positive rate. A random classifier would correspond to a 50% AUROC. Moreover, the metric sidesteps the issue of threshold selection.

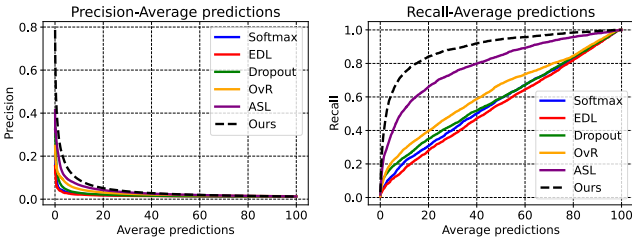
To be more clear, the confusion terms between all classes and the predicted class are regarded as the input to the metrics, while the target is one-hot class labels. As the ground-truth class could be imbalanced among misclassified samples, the AUROC is weighted by the categorical base rate.

The results on CIFAR-10, CIFAR-100 [26], and ImageNet [12] with 10, 100, and 1000 classes, respectively, are shown in Tab. 1 to demonstrate the comparison with different class scales. The Imagenet used here is an official down-sampled version with the resolution of 64×64. Besides standard Softmax, the Dropout [15] refers to the method using Monte Carlo Dropout (with the dropout rate of 0.2 and 10 dropout iterations) to approximate the distribution of model parameters. Since these methods do not contain explicit outputs of confusion, they are compromised to an intuitive approach, which measures the difference between predicted class probabilities. Two multi-label classification methods, namely, the One-vs-Rest (OvR) and the ASL [7], are included as baselines whose outputs are not normalized between classes. For EDL [44], confusion is defined as the difference between estimated singleton beliefs. In stark contrast to these methods, the proposed method uses explicitly estimated confusion between any classes with the predicted class following the definition in Eq. 5.

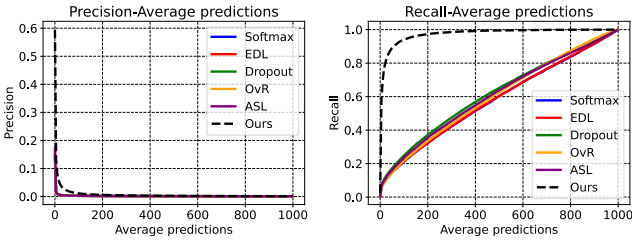
As shown in Tab. 1, the performance of the proposed method is significantly higher than all other approaches on three datasets. There are two reasons to support this result. First, the probability produced by softmax should not be viewed as the direct measure of confidence for each class [18], which means there might be no suitable way to derive confusion from softmax probability after training. Second, for the evidential method [44], the difference between predicted beliefs does not capture their shared features. In other words, the confusion in EDL is mixed with ignorance in their one uncertainty estimate. Orthogonal to



(a) Results on CIFAR-10



(b) Results on CIFAR-100



(c) Results on Imagenet

Figure 5: Precision and recall of delivering multiple predictions on misclassified samples with respect to the average number of predictions. The intention to predict extra classes is dependent on the confusion term between classes.

these methods, the proposed approach could model the confusion that occurred between classes and could help flexible recognition systems determine when and which class to be predicted next.

5.3. Flexible Closed-set Recognition

We turn to the task of flexible visual recognition on the closed CIFAR-10, CIFAR-100, and Imagenet, where methods are supposed to make multiple predictions when they are unsure. And we regard the output as correct as long as the correct answer is included in the prediction set. However, simply increasing the number of predictions for each sample should be penalized. Thus, for a more meaningful comparison, we show precision and recall as the function of the average number of predictions. We only evaluate methods on their misclassified samples, where providing more predictions is urgently demanded. And the indicator for making another prediction is the confusion between the considered and the predicted class, the same as in Section 5.2. The confusion for the first-predicted class itself is intuitively described as 0.

Closed Dataset CIFAR-10	+ LSUN (crop)	+ ImageNet (crop)
Softmax	64.2	63.9
OpenMax [8]	65.7	66.0
OSRCI [38]	65.0	63.6
LadderNet + OpenMax [52]	65.2	65.3
DHRNet + OpenMax [52]	65.6	65.5
CROSR [52]	72.0	72.1
GFROSR [41]	75.1	75.7
Ours	**80.5**	**76.8**

Table 2: Comparison on open-set detection by adding different unknown samples to the test set. The performance is evaluated by Macro-F1, which considers $K + 1$ classes.

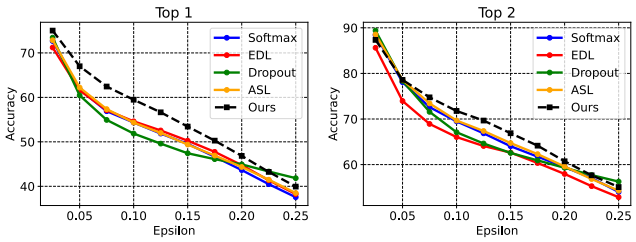


Figure 6: Top 1 and 2 accuracies with adversarial perturbations ϵ on CIFAR-10 [26] dataset.

The curves of three datasets are drawn in Fig. 5. We incrementally select more predictions for each class, *i.e.*, the samples are deserved to make an extra prediction if its confusion on this considered class is more significant than in other samples. By doing this, we could control the average predictions while being threshold-independent in evaluating flexible visual recognition. A desired precision-average prediction curve would have a near 1 value for the start of making predictions. For the recall-average prediction curve, the closer it adheres to the left-top corner, the earlier it delivers the correct class. When comparing the precision curve, our method could achieve a significantly higher value than other methods when making limited predictions. For example, on the CIFAR-10 dataset, the proposed method reaches the precision of 0.62 for making an average of only one prediction for each sample, while the second highest is 0.33. Note some samples could have no prediction at the beginning as all of their confusion terms are lower than other samples. Besides, we also notice that ASL [7] is better than other baselines in the recall curve. This indicates that for flexible recognition, a multi-label classifier could be a better choice as it prevents the overconfident problem in the softmax probabilities to some extent.

5.4. Ignorance for Open-set Detection

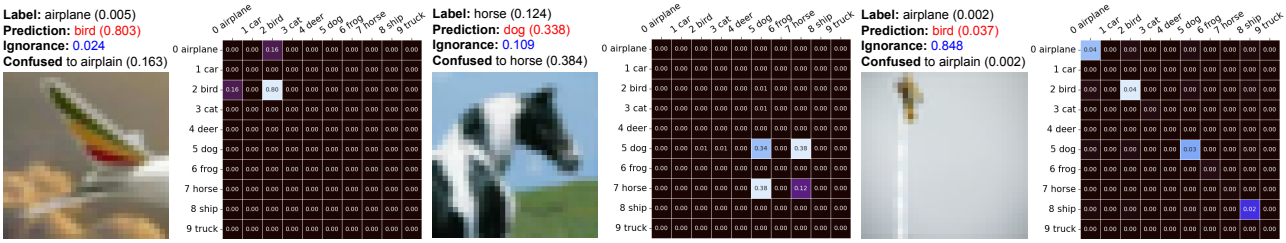
As a crucial part of flexible recognition, we additionally demonstrate the effectiveness of our method on the task of open-set detection in Tab. 2 following standard protocols [52, 18, 29]. The same network architecture, *i.e.*, a 13-layer VGG model, is used to implement the proposed method as in [52].

During the test time of open-set detection, the test images from CIFAR-10 datasets are viewed as closed-set examples. For open-set samples, we consider two natural image datasets, *i.e.*, LSUN (crop) and ImageNet (crop), introduced by Liang et al [29]. The Macro-F1 is therefore evaluated on $K + 1$ classes by regarding all open samples as an additional class. And the threshold for being detected as open-set is chosen when 95% of closed-set images can be correctly classified. Among the compared methods, OSRCI [38] augments the training set with hard and counterfactual generated images to improve unknown sample detection during testing. CROSR [52] and GFROSR [41] incorporate reconstruction loss into the procedure of closed-set training to better model the data distribution. For the proposed method, we use ignorance $\mathcal{I}$ as the indicator of unknown samples and achieve better performance without additional adversarial training data. The F1 score of the proposed method is higher than other methods on both datasets, which supports the claim that ignorance in the proposed method could effectively handle the lack of evidence in open-set samples. It is worth noting that, besides explicitly delivering ignorance to detect unknown samples, the advantage of the proposed method also lies in the estimate of confusion, which is evaluated in closed-set experiments.

5.5. Performance on Adversarial Samples

We compare the robustness of different classification methods [15, 44] against adversarial attacks. Besides models introduced in the previous experiment, we include the recent multi-label classification method ASL [7] as another baseline. The reason is that multi-label classification methods usually adopt the same sigmoid activation and multiple binary linear layers as our method. For our method, we use the plausibility pl_i for class i to predict instead of their singleton beliefs, *i.e.*, the confusion for each class that occurred during adding perturbations is involved in the ranking. The reason is that we want to test whether the confusion correctly characterizes the conflicting evidence shared between the correct and other classes.

Adversarial samples are generated on CIFAR-10 [26] using the Fast Gradient Sign method [17] with various perturbation parameters ϵ. The adversarial attack method uses the gradient during inference to generate samples that are more challenging to make correct predictions as ϵ increases. Both the top 1 and top 2 results are shown in Fig. 6. The figure indicates that the proposed method could almost achieve both the highest top 1 and top 2 results. That is, the confusion that happened during perturbating is captured by our method, which could still contribute to the corresponding correct class. The proposed method is only slightly worse than the Dropout method when $\epsilon = 0.25$, which is forgivable as its training includes stochastic zeroing on the model parameter to increase its robustness.



(a) (b) (c)

Figure 7: Matrices of confusion of misclassified samples on the CIFAR-10 dataset. The diagonal of each matrix is set to the singleton belief of each class. Notice that the background color for each heatmap is normalized individually. The total ignorance is demonstrated in the caption.

Method	CIFAR-10 + LSUN (crop)				
	Acc.	F1	AUROC	AUPR closed	AUPR open
Ours w/o Reg.	72.4	84.8	96.2	96.9	95.3
Ours w/ Reg.	**79.9**	**86.9**	**97.0**	**97.5**	**96.6**
	CIFAR-10 + ImageNet (crop)				
Ours w/o Reg.	63.9	82.0	95.1	96.1	93.8
Ours w/ Reg.	**71.7**	**83.9**	**95.8**	**96.5**	**94.7**

Table 3: Ablation studies about the regularization term $\mathcal{L}_{\text{reg}}$ on the task of open-set detection with two hybrid datasets.

5.6. Qualitative Results

Furthermore, to provide a more intuitive impression of our results, we demonstrate the matrics of binary confusion estimates in Fig. 7. We take the misclassified images of the proposed method for the CIFAR-10 dataset. The ground truth label, the prediction, the singleton belief, the ignorance, and the highest binary confusion are noted in their captions. Each grid in the confusion matrix denotes the confusion between any two classes, while the diagonal represents the singleton belief.

In Fig. 7b, the proposed method wrongly recognizes the image into *dog* class while maintaining a very high confusion towards the correct class *horse*. However, as the singleton belief is not significant for the predicted class, the proposed method could predict a combined result, *i.e.*, both *dog* and *horse*, if setting a relatively high threshold on the sum of beliefs. Another example in Fig. 7c implies the ability of our method to reject making a prediction. As the image in Fig. 7c is contentless and low-resolution, the proposed failed to collect enough supportive evidence for any known classes, which delivers a very high ignorance value.

5.7. Ablation Studies

We investigate how the regularization term $\mathcal{L}_{\text{reg}}$ influences performance. The reason we add the regularization is to avoid the first output of each plausibility function converging to the belief instead of the plausibility. To show its effectiveness, we evaluate the proposed method on the task of open-set detection in Tab. 3. The separation could be deemed to be more sufficient if the ignorance term performs better in detecting open-set samples. Note we use the ResNet-18 as our backbone here to remain consistent with our flexible recognition experiments. For the metric of open-set detection accuracy and AUROC, closed-set samples are regarded as negative samples, while open-set samples constitute positive samples. The other two metrics, AUPR closed and AUPR open, denote the Area Under the Precision-Recall curve where closed-set or open-set images are specified as positives, respectively. The overall improvements in adding the regularization term $\mathcal{L}_{\text{reg}}$, as demonstrated in Tab. 3, indicate its effectiveness in promoting the separation.

5.8. Discussions and Future works

In our experiments, we find the scale of confusion and ignorance varies with different backbones and datasets. The correlation behind it could be related to the capabilities of different models and the learning difficulties of different datasets. We leave the investigation for our future work.

6. Conclusions

In this paper, we introduce a novel approach to explicitly model two distinct sources of uncertainties, *i.e.*, confusion and ignorance, under the novel task of flexible recognition. The recognition system is expected to reject samples from unknown classes and also make multiple predictions when it is uncertain about a closed-set image. Particularly, in the proposed method, the confusion is modeled as the conflicting evidence, while the ignorance represents the total lack of evidence. The hypothesis space of recognition is then divided and modeled by multiple plausibility functions. The model learns the concentration parameter of Dirichlet prior, which is being placed on the belief for singletons. A complete set of opinions could be generated through evidence combinations. Experiments on different datasets, along with challenging tasks of adversarial disturbance, flexible recognition, and open-set detection, confirm the effectiveness of the proposed method.

References

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020. 1, 2

[2] Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020. 2

[3] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. 2

[4] Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*, 2020. 2

[5] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13349–13358, 2021. 2

[6] Jeffrey A Barnett. Computational methods for a mathematical theory of evidence. In *Classic Works of the Dempster-Shafer Theory of Belief Functions*, pages 197–216. Springer, 2008. 2, 3

[7] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. *arXiv preprint arXiv:2009.14119*, 2020. 6, 7

[8] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016. 2, 7

[9] Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in Neural Information Processing Systems*, 33:1356–1367, 2020. 1

[10] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III*, pages 507–522, 2020. 2

[11] Charles Corbière, Marc Lafon, Nicolas Thome, Matthieu Cord, and Patrick Pérez. Beyond first-order uncertainty estimation with evidential models for open-world recognition. In *ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning*, 2021. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[13] Nemanja Djuric, Vladan Radosavljevic, Henggang Cui, Thi Nguyen, Fang-Chieh Chou, Tsung-Han Lin, Nitin Singh, and Jeff Schneider. Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2095–2104, 2020. 1

[14] Michael Dusenberry, Ghassen Jerfel, Yeming Wen, Yian Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable bayesian neural nets with rank-1 factors. In *International conference on machine learning*, pages 2782–2792. PMLR, 2020. 2

[15] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015. 1, 2, 6, 7

[16] ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*, 2017. 2

[17] Ian Goodfellow, Nicolas Papernot, Patrick McDaniel, Reuben Feinman, Fartash Faghri, Alexander Matyasko, Karen Hambardzumyan, Yi-Lin Juang, Alexey Kurakin, Ryan Sheatsley, et al. cleverhans v0. 1: an adversarial machine learning library. *arXiv preprint arXiv:1610.00768*, 1, 2016. 7

[18] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 6, 7

[19] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017. 2, 3

[20] Stephen C Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2-3):217–223, 1996. 2

[21] Audun Jøsang. *Subjective logic*, volume 3. Springer, 2016. 1, 2, 3

[22] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015. 2

[23] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 2

[24] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2021. 2

[25] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pages 5436–5446. PMLR, 2020. 2

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6, 7

[27] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, 142:106816, 2020. 2

[28] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 1, 2

[29] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 7

[30] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 2

[31] Bo Liu, Hao Kang, Haoxiang Li, Gang Hua, and Nuno Vasconcelos. Few-shot open-set recognition using meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2020. 2

[32] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020. 1

[33] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018. 1

[34] Andrey Malinin and Mark Gales. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[35] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27, 2014. 1

[36] Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12):3868–3878, 2020. 1

[37] Rhiannon Michelmore, Marta Kwiatkowska, and Yarin Gal. Evaluating uncertainty quantification in end-to-end autonomous driving control. *arXiv preprint arXiv:1811.06817*, 2018. 1

[38] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 613–628, 2018. 2, 7

[39] Poojan Oza and Vishal M Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2307–2316, 2019. 2

[40] Badri N Patro, Mayank Lunayach, Shivansh Patel, and Vinay P Namboodiri. U-cam: Visual explanation using uncertainty based class activation maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7444–7453, 2019. 1

[41] Pramuditha Perera, Vlad I Morariu, Rajiv Jain, Varun Manjunatha, Curtis Wigington, Vicente Ordonez, and Vishal M Patel. Generative-discriminative feature representations for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11814–11823, 2020. 2, 7

[42] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8491–8501. PMLR, 13–18 Jul 2020. 2

[43] Murat Sensoy, Lance Kaplan, Federico Cerutti, and Maryam Saleki. Uncertainty-aware deep classifiers using generative models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34,04, pages 5620–5627, 2020. 1

[44] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018. 1, 2, 3, 4, 5, 6, 7

[45] Rui Shao, Pramuditha Perera, Pong C Yuen, and Vishal M Patel. Open-set adversarial defense. In *European Conference on Computer Vision*, pages 682–698. Springer, 2020. 2

[46] Xin Sun, Zhenning Yang, Chi Zhang, Keck-Voon Ling, and Guohao Peng. Conditional gaussian distribution learning for open set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13480–13489, 2020. 2

[47] Dustin Tran, Mike Dusenberry, Mark van der Wilk, and Danijar Hafner. Bayesian layers: A module for neural network uncertainty. *Advances in neural information processing systems*, 32, 2019. 2

[48] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005. 2

[49] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019. 1, 2

[50] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4921–4930, 2022. 2

[51] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. *Advances in Neural Information Processing Systems*, 33:6514–6527, 2020. 2

[52] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4016–4025, 2019. 2, 7

[53] Kuang Zhou, Arnaud Martin, and Quan Pan. Evidence combination for a large number of sources. In *2017 20th International Conference on Information Fusion (Fusion)*, pages 1–8. IEEE, 2017. 1