Weakly-Supervised Residual Evidential Learning for Multi-Instance Uncertainty Estimation

Pei Liu [1] **Luping Ji** [1]

Abstract

Uncertainty estimation (UE), as an effective means to quantify predictive uncertainty, is crucial for safe and reliable decision-making, especially in high-risk scenarios. Existing UE schemes usually assume that there are completely-labeled samples to support fully-supervised learning. In practice, however, many UE tasks often have no sufficiently-labeled data to use, such as the Multiple Instance Learning (MIL) with only weak instance annotations. To bridge this gap, this paper, for the first time, addresses the weakly-supervised issue of *Multi-Instance UE* (MIUE) and proposes a new baseline scheme, *Multi-Instance Residual Evidential Learning* (MIREL). Particularly, at the fine-grained instance UE with only weak supervision, we derive a multi-instance residual operator through the Fundamental Theorem of Symmetric Functions. On this operator derivation, we further propose MIREL to jointly model the high-order predictive distribution at bag and instance levels for MIUE. Extensive experiments empirically demonstrate that our MIREL not only could often make existing MIL networks perform better in MIUE, but also could surpass representative UE methods by large margins, especially in instance-level UE tasks. Our source code is available at https://github.com/liupei101/MIREL.

1. Introduction

Deep learning models have shown impressive capability and become ubiquitous in the last decade. However, they often tend to produce overconfident predictions, even for shifted or unseen samples (Nguyen et al., 2015; Kendall & Gal, 2017). Such behaviour may lead to disastrous consequences in safety-critical scenarios, *e.g.*, autonomous driving and medical diagnosis (Franchi et al., 2022; Linmans et al., 2023), calling into question their real-world usability. Thus, it is particularly important to provide an accurate confidence level for the prediction of neural networks (NNs) through uncertainty estimation (UE) methods.

To accomplish accurate UE, epistemic (model) uncertainty is considered generally (Postels et al., 2022; Mukhoti et al., 2023). From a principled Bayesian perspective (Neal, 2012), it is characterized by the distribution over model parameters given training data $\mathcal{D}$, *i.e.*, $p(\boldsymbol{\omega}|\mathcal{D})$. Due to its involvement, a model parameter $\boldsymbol{\omega}$ would be low in likelihood when it is incompatible with $\mathcal{D}$; as a result, less confidence would be yielded for the new samples shifted in distribution, thus alleviating overconfidence in prediction. However, most current practices often assume that there are *completely-labeled* samples in $\mathcal{D}$ with which NNs can be trained to support $p(\boldsymbol{\omega}|\mathcal{D})$ (Mena et al., 2021).

In fact, there are many practical tasks involving *weakly-annotated* data, in which no complete label can be directly utilized for training. While such tasks remain under-explored in UE, a fundamental machine learning problem we consider under this class is multiple instance learning (MIL) (Dietterich et al., 1997). As a typical task of weakly-supervised learning, it is prominent in many labeling-intensive applications, *e.g.*, histopathology diagnosis (Ilse et al., 2020; Liu et al., 2024a), video anomaly detection (Sultani et al., 2018; Zhong et al., 2019) and video analysis (Babenko et al., 2010; Rizve et al., 2023), etc. Specifically, in histopathology diagnosis an image usually contains gigapixels, so it is often divided into thousands of small patches for MIL, where multiple patches (instances) are observed but only a general statement of their labels is given. In this case, a diagnosis model has to learn from weakly-annotated patches to make patch-level predictions. UE is highly anticipated to provide accurate uncertainty measures for these weakly-supervised predictions to make final diagnostic decisions safer and more reliable.

Generally, a sample given in MIL is described as a bag X and its label is known, $Y \in \{0, 1\}$. In particular, X is composed of multiple instances, *i.e.*, $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$, but instance labels $\{y_1, \ldots, y_K\}$ are *unknown* and instead

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. Correspondence to: Luping Ji <jiluping@uestc.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

connect with bag label via a classical MIL assumption, $Y = \max_k\{y_k\}$ and $k \in [1, K]$. Under these special settings, MIL is usually interested in two tasks, bag-level and instance-level prediction (Kandemir & Hamprecht, 2015). This means that, a MIL model needs to **i**) learn $p(\boldsymbol{\omega}|\mathcal{D})$ from the multi-instance bags with *variable* sizes and meanwhile **ii**) jointly estimate a new *weakly-supervised posterior* $p(\boldsymbol{\theta}_{\mathbf{w}}|\mathcal{D})$ from weakly-annotated instances. Therefore, in such MIL tasks, the common way of capturing epistemic uncertainty for UE seems less practical. This motivates us to focus on the problem of Multi-Instance UE (termed **MIUE**) and study a baseline approach for it.

Bag-level The Fundamental Theorem of Symmetric Functions (Zaheer et al., 2017) provides a general strategy to score a bag of instances. It deals with size-varied bags by a permutation-invariant MIL pooling operator (Ilse et al., 2018). Accordingly, $p(\boldsymbol{\omega}|\mathcal{D})$ could be estimated from fully-labeled bags using common UE techniques to capture bag-level epistemic uncertainty, like that in fully-supervised learning (Mena et al., 2021).

Instance-level Without complete instance labels, modeling predictive uncertainty at instance level would not be as straightforward as that at bag level. Nonetheless, attention-based MIL (Ilse et al., 2018; Li et al., 2021) still could generate instance predictions with their instance scoring proxy—*attention branch*. Following this approach, various strategies (Shi et al., 2020; Qu et al., 2022; Cui et al., 2023) are proposed to make instance prediction more accurate. We argue that such attention-dependent means may not be suitable for learning $p(\boldsymbol{\theta}_{\mathbf{w}}|\mathcal{D})$ jointly with $p(\boldsymbol{\omega}|\mathcal{D})$. Because i) the parameter $\boldsymbol{\theta}_{\mathbf{w}}$ given by attention branch for instances is completely contained within the parameter $\boldsymbol{\omega}$ for bags, and notably ii) the attention scores given by that proxy are produced essentially for learning better bag representations, leaving a substantial gap to ideal instance predictions.

In this paper, with the Fundamental Theorem of Symmetric Functions, we demonstrate that the gap to ideal instance prediction can be narrowed by turning to exploit a good bag-level decision space. With this basic finding, we propose a new MIL scheme for MIUE. Concretely, (1) we devise a new instance estimator to jointly learn $p(\boldsymbol{\theta}_{\mathbf{w}}|\mathcal{D})$ by deriving a *multi-instance residual operator*. This operator makes instance prediction separated from bag decision. (2) Further, we model high-order probability distribution at bag and instance levels to fulfill MIUE, by parameterizing two Dirichlet distributions with the evidences provided by general bag estimator and our residual instance estimator. (3) To optimize $\boldsymbol{\theta}_{\mathbf{w}}$ without complete instance labels, we propose a weakly-supervised evidence learning strategy and prove that it provides a tighter upper bound for ideal instance loss function than common strategies under given conditions.

The main contributions of this paper are summarized as follows: (1) A new problem of uncertainty estimation, termed MIUE (*Multi-Instance UE*), is introduced in this paper. To our knowledge, we are the first to study it in MIL. (2) With the Fundamental Theorem of Symmetric Functions, this paper demonstrates that a good estimator for instances can be directly deduced from a good bag-level decision space, no longer relying on the scoring proxy from attention-based MIL. (3) This paper further derives a residual estimator specially for instances, and proposes a new scheme, *Multi-Instance Residual Evidential Learning* (MIREL), for MIUE. This scheme can jointly quantify the predictive uncertainty at bag and instance levels in MIL.

2. Preliminary

2.1. Multiple Instance Learning (MIL)

Definition Here we give the formal conventions and notations in MIL, following Ilse et al. (2018). A given sample (bag) is denoted as $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$, where $\mathbf{x}_1, \ldots, \mathbf{x}_K$ are usually treated to be *i.i.d.* Its label, $Y \in \{0, 1\}$, is accessible for training; its instance-level labels, $\{y_1, \ldots, y_K\}$, are *unknown* and $y_k \in \{0, 1\}$ for $k \in [1, K]$. A classical MIL assumption states that, a bag is positive ($Y = 1$) *iff* it has at least one positive instance; otherwise, it is negative ($Y = 0$). Namely, there is $Y = \max_k\{y_k\}$.

Learning paradigm To learn from size-varied bags, a common practice is to leverage a *permutation-invariant* pooling operator. It can be expressed by the Fundamental Theorem of Symmetric Functions as follows:

Theorem 2.1 ((Zaheer et al., 2017; Ilse et al., 2018))**.** *A scoring function for a set of instances $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$, $S(X) \in \mathbb{R}$, is a symmetric function (permutation-invariant to the elements in X), if and only if it can be written as*

$$S(X) = g\Big(\sum_{k=1}^{K} f(\mathbf{x}_k)\Big), \tag{1}$$

where f and g are suitable transformations.

This theorem holds as before or under weak conditions, when the form of instance pooling in Eq.(1), $\sum_k f(\mathbf{x}_k)$, is replaced by others, such as i) mean, ii) $\max$ (Qi et al., 2017), and iii) attention-based MIL pooling.

Attention-based MIL pooling The most representative one is ABMIL (Ilse et al., 2018). It first proposes to leverage dynamic instance weights for pooling, written as $\sum_{k=1}^{K} a_k f(\mathbf{x}_k)$, where a_k is called attention score and

$$a_k = \mathrm{softmax}\left(t(\mathbf{h}_k)\right) = \frac{\exp\left(t(\mathbf{h}_k)\right)}{\sum_{\tau=1}^{K} \exp\left(t(\mathbf{h}_\tau)\right)}. \tag{2}$$

$\mathbf{h}_k$ stands for the instance embedding given by $\mathbf{h}_k = f(\mathbf{x}_k)$ and $t(\cdot)$ is a transformation parameterized by NNs.

Owing to attention mechanism, most attention-based MIL networks can provide a *proxy* (*i.e.*, attention score a_k) to estimate instance labels, as highlighted in Ilse et al. (2018) and Li et al. (2021). This proxy is frequently used and improved afterwards (Qu et al., 2022; Cui et al., 2023), often taken as a reliable estimator for instances.

2.2. Evidential Deep Learning (EDL)

As one of general UE methods, recently-proposed EDL (Sensoy et al., 2018) models predictive uncertainty using the Dempster–Shafer Theory of Evidence (DST) (Dempster, 1968). It formalizes the belief assignment in DST with Subjective Logic (SL) (Jøsang, 2016).

Belief assignment Considering C ($C \geq 2$) mutually exclusive singletons (*e.g.*, class labels), SL assigns a belief mass b_i to the i-th singleton for $i \in [1, C]$ and defines an overall uncertain mass u. Let $b_i = \frac{e_i}{\sum_{i=1}^{C}(e_i+1)} \geq 0$ and $u = \frac{C}{\sum_{i=1}^{C}(e_i+1)} \geq 0$, where e_i is the evidence of the i-th singleton. There is $u + \sum_{i=1}^{C} b_i = 1$, *i.e.*, a weaker belief over singletons indicates a higher overall uncertainty.

Posterior Dirichlet distribution SL further formalizes the belief assignment stated above as a Dirichlet distribution, offering the potential approach to modeling predictive probability. Specifically, let $Dir(\boldsymbol{p}|\boldsymbol{\alpha})$ denote a Dirichlet distribution, where $\boldsymbol{p} \in \mathcal{S}^{C-1}$ (a probability simplex with $C-1$ dimensions), $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_C]$, and $\alpha_i \geq 0$. By definition, there are $Dir(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{C} p_i^{\alpha_i - 1}$ and $B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{C}\Gamma(\alpha_i)}{\Gamma(\alpha_0)}$, where $B(\boldsymbol{\alpha})$ is a multinomial Beta function, $\Gamma(\cdot)$ is a *gamma* function, and $\alpha_0 = \sum_{i=1}^{C} \alpha_i$ often called the precision or Dirichlet strength. Therefore, predictive probability can be expressed by a posterior Dirichlet distribution $Dir(\boldsymbol{p}|\boldsymbol{\alpha})$, by deriving evidences $\boldsymbol{e}$ for C singletons and then adopting $\boldsymbol{\alpha} = \boldsymbol{e} + 1$ to parameterize $Dir(\boldsymbol{p}|\boldsymbol{\alpha})$, where $\boldsymbol{e} = [e_1, \cdots, e_C]$.

Evidential learning For any sample X, EDL employs a NN-based transformation Φ to derive evidences, *i.e.*, $\boldsymbol{e} = \Phi(X)$, forming a belief assignment over C classes. This assignment is then formalized as $Dir(\boldsymbol{p}|\boldsymbol{\alpha})$ with $\boldsymbol{\alpha} = \Phi(X) + 1$, to obtain the prediction of X. Compared with the standard neural classifiers that predict a Categorical distribution over classes, EDL predicts *a distribution over Categorical distribution*, the conjugate prior of Categorical distribution, thus modeling second-order probability distribution for UE (Jøsang, 2016; Sensoy et al., 2018).

By using NNs to parameterize $Dir(\boldsymbol{p}|\boldsymbol{\alpha})$, EDL provides a simple yet efficient *deterministic* method for UE. It can distinguish different uncertainties originated from data, model, or distribution (Ulmer et al., 2023). Moreover, Deng et al. (2023) show that EDL can be cast as learning PAC-Bayesian generalization bounds. These favorable advantages motivate us to study a baseline approach for MIUE through EDL.

3. Problem Formulation: Multi-Instance Uncertainty Estimation (MIUE)

First of all, we formalize MIUE from the perspective of Bayesian (Neal, 2012), as it offers a principled way to study predictive uncertainty and is widely adopted in UE.

Bag-level Given a *bag dataset* $\mathcal{D} = \{(X_j, Y_j)\}_{j=1}^{N}$ and the parameter $\boldsymbol{\omega}$ of any MIL model, the prediction of a bag X^* can be written as a posterior distribution:

$$P(Y^*|X^*, \mathcal{D}) = \int P(Y^*|X^*, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\mathcal{D}) d\boldsymbol{\omega}. \quad (3)$$

This predictive form reflects that the uncertainty in bag prediction results from data (aleatoric) and model (epistemic) uncertainty (Der Kiureghian & Ditlevsen, 2009; Kendall & Gal, 2017), captured by $P(Y^*|X^*, \boldsymbol{\omega})$ and $p(\boldsymbol{\omega}|\mathcal{D})$, respectively. $\boldsymbol{\omega}$ parameterizes the mapping from X to Y.

Instance-level For any $(X_j, Y_j) \in \mathcal{D}$, $X_j = \{\mathbf{x}_{jk}\}_{k=1}^{K_j}$, where K_j is the instance number of X_j. Following the assumption of MIL, there is $Y_j = \max\{y_{jk}\}_{k=1}^{K_j}$ and y_{jk} is *unknown*. The prediction of an instance $\mathbf{x}^*$ is expressed as

$$P(y^*|\mathbf{x}^*, \mathcal{D}) = \int P(y^*|\mathbf{x}^*, \boldsymbol{\theta}_{\mathbf{w}}) p(\boldsymbol{\theta}_{\mathbf{w}}|\mathcal{D}) d\boldsymbol{\theta}_{\mathbf{w}}, \quad (4)$$

where $\boldsymbol{\theta}_{\mathbf{w}}$ is the parameter estimated from *weakly-annotated* instances to represent the mapping from $\mathbf{x}$ to y. $\boldsymbol{\theta}_{\mathbf{w}}$ and $\boldsymbol{\omega}$ could share some parameters in their joint learning on $\mathcal{D}$.

Uncertainty measures To quantify the uncertainty in prediction, various measures could be adopted (Malinin & Gales, 2018). *Max probability* and the *entropy* of expected predictive distribution are the most frequently used two for measuring total uncertainty. Moreover, *expected entropy* and *mutual information* (MI) are often adopted as the measures to capture data uncertainty and model uncertainty, respectively. More details are elaborated in Appendix B.

Traditional non-Bayesian MIL networks usually treat $\boldsymbol{\omega}$ and $\boldsymbol{\theta}_{\mathbf{w}}$ deterministically, ignoring the model uncertainty in bag and instance predictions. This could lead to failures on out-of-distribution samples (Blanchard et al., 2011). By contrast, Bayesian frameworks account for both data and model uncertainties, allowing us to quantify multi-instance uncertainty more accurately and holistically.

4. Method

4.1. Bag-level Predictive Uncertainty

Bag-level predictive uncertainty can be quantified via Eq.(3). However, integrating over a high-dimensional space of $\boldsymbol{\omega}$

is often *intractable*. To tackle this, we propose to model bag-level predictive probability with a posterior Dirichlet distribution, inspired by EDL.

Evidential learning Concretely, for a bag $X = \{\mathbf{x}_k\}_{k=1}^K$, we express its predictive probability as follows:

$$\begin{aligned} p(\boldsymbol{\mu}|X) &= Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}_{\text{bag}}), \\ \boldsymbol{\alpha}_{\text{bag}} = \boldsymbol{e}_{\text{bag}} + 1 &= \mathcal{A}(S(X)) + 1 \\ &= \mathcal{A}\Big(g_\phi\big(\sum_k f_\psi(\mathbf{x}_k)\big)\Big) + 1, \end{aligned} \tag{5}$$

where $\boldsymbol{\alpha}_{\text{bag}}$ is the concentration parameter of bag-level Dirichlet distribution, $\boldsymbol{e}_{\text{bag}}$ is the bag evidence collected from X, $\mathcal{A}(\cdot)$ is an activation function for non-negative evidence outputs, and the transformations f and g are parameterized by the NNs with parameters ψ and ϕ, respectively. To optimize parameters, we adopt a Fisher Information-based objective function (Deng et al., 2023), proven to be well-suited for EDL. It is denoted by $\mathcal{L}_{\mathcal{I}\text{-EDL}}$ as follows:

$$\min_{\psi,\phi} \mathbb{E}_{(X,Y)\sim\mathcal{P}} \mathbb{E}_{\boldsymbol{\mu}\sim Dir(\boldsymbol{\alpha}_{\text{bag}})}\big[-\log p(Y|\boldsymbol{\mu}, \boldsymbol{\alpha}_{\text{bag}}, \sigma^2)\big], \tag{6}$$

where $p(Y|\boldsymbol{\mu}, \boldsymbol{\alpha}_{\text{bag}}, \sigma^2)$ is assumed to be a multivariate Gaussian distribution $\mathcal{N}\big(Y|\boldsymbol{\mu}, \sigma^2\mathcal{I}(\boldsymbol{\alpha}_{\text{bag}})^{-1}\big)$ and $\mathcal{I}(\boldsymbol{\alpha}_{\text{bag}})$ is the Fisher Information Matrix of $Dir(\boldsymbol{\alpha}_{\text{bag}})$. We give its details in Appendix C for completeness.

Justification (1) $p(\boldsymbol{\mu}|X)$: the predictive form of bag X given in Eq.(5), *i.e.*, a posterior Dirichlet distribution rather than a conventional Categorical one, can be derived from Eq.(3) approximately (Malinin & Gales, 2018), *i.e.*,

$$\begin{aligned} \int P(Y|X, \boldsymbol{\omega})p(\boldsymbol{\omega}|\mathcal{D})d\boldsymbol{\omega} &= \int P(Y|\boldsymbol{\mu})p(\boldsymbol{\mu}|X, \mathcal{D})d\boldsymbol{\mu} \\ &\approx \int P(Y|\boldsymbol{\mu})p(\boldsymbol{\mu}|X; \hat{\boldsymbol{\omega}})d\boldsymbol{\mu}, \end{aligned} \tag{7}$$

with a point estimation $\hat{\boldsymbol{\omega}}$ satisfying $p(\boldsymbol{\omega}|\mathcal{D}) = \delta(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}})$ where $\delta(\cdot)$ is a Dirac delta function. Eq.(7) makes the full posterior in Eq.(3) tractable by introducing a new posterior Dirichlet distribution, thus enabling us to quantify the uncertainty in $p(\boldsymbol{\mu}|X)$ deterministically. More details are provided in Appendix A.1. **(2)** $\boldsymbol{\alpha}_{\textbf{bag}}$: its evidential learning formulation in Eq.(5) still satisfies the condition of Theorem 2.1. This thereby provides broad MIL networks with an alternative means to model bag-level uncertainty.

4.2. Rethinking Instance-level Estimator

It is often of interest to ask MIL models to jointly estimate instance labels. For this purpose, herein, we first consider the estimator for instances before modeling instance-level predictive uncertainty. Instead of scoring instances with attention mechanism as written in Eq.(2), we rethink and derive a new estimator for instance prediction through the Fundamental Theorem of Symmetric Functions.

Given the unified form of permutation-invariant bag classifiers (Theorem 2.1), *i.e.*, $S(X) = g\big(\sum_k f(\mathbf{x}_k)\big)$, it could be found that a single aggregated embedding (by pooling over instance embeddings) is transformed into a bag score by g. Obviously, it is also possible to transform a single instance embedding for scoring, thereby obtaining a *feasible* estimator for instances, as summarized below:

Corollary 4.1. *Given a scoring function for a set of instances $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$, written as $S(X) = g\big(\sum_k f(\mathbf{x}_k)\big) \in \mathbb{R}$ where $k \in [1, K]$, a scoring function for any single instance can be written as*

$$T = g \circ f, \tag{8}$$

and $T(\mathbf{x}) \in \mathbb{R}$ for an instance $\mathbf{x}$.

Its proof is shown in Appendix A.2. This corollary allows us to make instance prediction by *skipping* permutation-invariant pooling. Nonetheless, it is still insufficient to conclude that T is also a good instance estimator, apart from a feasible one. We thereby give Proposition 4.2.

Proposition 4.2. *Let $S(\cdot)$ be a classifier for a bag of instances $X = \{\mathbf{x}_k\}_{k=1}^K$ and satisfy $S(X) = g\big(\sum_k f(\mathbf{x}_k)\big)$. For any bag X and its label $Y \in \{0, 1\}$, further assume S can predict bags precisely: $S(X) = Y$. Then, there exists an estimator with $T = g \circ f$ for any single instance $\mathbf{x}$, such that $T(\mathbf{x}) = y$, where $y \in \{0, 1\}$ is the label of $\mathbf{x}$.*

This proposition implies that a perfect T can be directly deduced from a perfect S. Refer to Appendix A.2 for its proof. Furthermore, by relaxing the ideal assumption, *i.e.*, a perfect S, it suggests that a good instance-level estimator is likely to be obtained from a good bag classifier. Intuitively, if S is good at distinguishing between positive and negative bags, it would recognize an instance $\mathbf{x}$ correctly as long as one *duplicates* $\mathbf{x}$ to form a new bag for prediction, because this bag is expected to be classified correctly by S and it exactly has the *same* label with $\mathbf{x}$. Extensive experiments (Section 6) empirically demonstrate this finding.

Compared with the commonly-used attention score a_k, our new instance-level estimator $T(\mathbf{x})$ exhibits the following merits. (1) It could be obtained from not merely attention-based but general MIL approaches, as stated in Corollary 4.1. (2) It is no longer a scoring proxy like a_k but an estimator with a classification decision space same to that of $S(X)$, probably closer to an ideal instance estimator than a_k.

4.3. Weakly-supervised Instance-level Predictive Uncertainty

Given $T(\mathbf{x}) = g_\phi(f_\psi(\mathbf{x}))$, T could be adopted as the estimator to infer instance probabilities, with the same f_ψ

and g_ϕ as S. However, it could be not desirable for UE in practice, because the NN-based g_ϕ is actually trained with those *pooled* instance embeddings (*i.e.*, $\sum_k f_\psi(\mathbf{x}_k)$) instead of raw ones (*i.e.*, $f_\psi(\mathbf{x}_k)$), thus inclining to estimate high uncertainties for *unseen* data. To this end, we propose to improve T by learning instance-specific residuals, thereby obtaining a new estimator specially for instance-level UE. Moreover, we propose a weakly-supervised evidence learning objective to optimize our residual instance estimator.

Residual evidential modeling For any single instance $\mathbf{x}$, we assume there is a *residual estimation* for the given T, denoted as $\epsilon = T^*(\mathbf{x}) - T(\mathbf{x})$ where $T^*(\mathbf{x})$ is a target estimation. We model this instance-specific residual ϵ by leveraging a new NN-parameterized transformation $r_\pi(\cdot)$ whose input is $\mathbf{h} = f_\psi(\mathbf{x})$. It is encouraged to compensate for the initial biased $T(\mathbf{x})$, so as to approach $T^*(\mathbf{x})$. Let $R(\cdot)$ denote our new residual instance estimator. Further, we introduce a Dirichlet distribution into R for instance-level UE. Therefore, the predictive distribution of instance $\mathbf{x}$ can be summarized and written as follows:

$$\begin{aligned} p(\boldsymbol{\nu}|\mathbf{x}) &= Dir(\boldsymbol{\nu}|\boldsymbol{\alpha}_{\text{ins}}), \\ \boldsymbol{\alpha}_{\text{ins}} &= \boldsymbol{e}_{\text{ins}} + 1 = \mathcal{A}\big(R(\mathbf{x})\big) + 1, \\ R(\mathbf{x}) &= T(\mathbf{x}) + r_\pi(\mathbf{h}) = g_\phi(f_\psi(\mathbf{x})) + r_\pi(f_\psi(\mathbf{x})), \end{aligned} \tag{9}$$

where $\boldsymbol{\alpha}_{\text{ins}}$ is instance-level concentration parameter and $\boldsymbol{e}_{\text{ins}}$ indicates the evidence derived from $\mathbf{x}$.

Optimization strategy Given any bag sample (X, Y) where $X = \{\mathbf{x}_k\}_{k=1}^K$, our optimization strategy for $R(\mathbf{x})$ is as follows. (1) $Y = 0$: we have $y_k = 0\ \forall k \in [1, K]$, so the objective function given by Eq.(6) can be utilized similarly. (2) $Y = 1$: since only $\max_k\{y_k\} = 1$ is given, we propose a weakly-supervised evidential learning strategy. Concretely, we simply multiply $\boldsymbol{e}_k$ (the evidence of $\mathbf{x}_k$) by different weights to mimic the selection of positive instances, and then aggregate them into a single one that can be optimized by Eq.(6). Those weights are the expected instance probabilities given by T, likely to indicate positive instances as stated in Section 4.2. Therefore, our objective function for $R(\mathbf{x})$ can be summarized and given as follows:

$$\begin{aligned} \mathcal{L}_{\text{MIREL}} &= \min_{\psi,\pi} \mathbb{E}_{(X,Y)\sim\mathcal{P}}\big[Y\mathcal{L}^+_{\text{ins}} + (1-Y)\mathcal{L}^-_{\text{ins}}\big], \\ \mathcal{L}^-_{\text{ins}} &= \mathbb{E}_{\boldsymbol{\nu}\sim Dir(\boldsymbol{\alpha}_k)}\big[-\log p(y=0|\boldsymbol{\nu}, \boldsymbol{\alpha}_k, \sigma^2)\big], \\ \mathcal{L}^+_{\text{ins}} &= \mathbb{E}_{\boldsymbol{\nu}\sim Dir(\widetilde{\boldsymbol{\alpha}})}\big[-\log p(Y=1|\boldsymbol{\nu}, \widetilde{\boldsymbol{\alpha}}, \sigma^2)\big], \end{aligned} \tag{10}$$

where $\boldsymbol{\alpha}_k$ is the $\boldsymbol{\alpha}_{\text{ins}}$ for $\mathbf{x}_k$, $\widetilde{\boldsymbol{\alpha}} = \sum_k \boldsymbol{e}_k \bar{w}_k + 1$, $\bar{w}_k$ is a normalized w_k, $w_k = \mathbb{E}_{\boldsymbol{\nu}\sim Dir\left(\boldsymbol{\alpha}_k^{(T)}\right)} p(y=1|\boldsymbol{\nu})$, and $\boldsymbol{\alpha}_k^{(T)} = \mathcal{A}(T(\mathbf{x}_k)) + 1$. Note that only ψ and π are optimized in Eq.(10), without the ϕ for bags, as we aim at encouraging $r_\pi \circ f_\psi$ to learn residuals for instances.

Justification (1) $R(\mathbf{x})$: this new residual instance estimator has independent parameters π with $S(X)$, as written in Eq.(9). This enables R to separately learn instance-specific evidences and enhance instance-level UE. **(2)** $\mathcal{L}^+_{\text{ins}}$: we analyze its upper bounds, given in Proposition 4.3. This proposition suggests that $\mathcal{L}^+_{\text{ins}}$ may provide a more suitable $\hat{\boldsymbol{\theta}}_{\mathbf{w}}$ than common objectives such that there is $p(\boldsymbol{\theta}_{\mathbf{w}}|\mathcal{D}) \approx \delta(\boldsymbol{\theta}_{\mathbf{w}} - \hat{\boldsymbol{\theta}}_{\mathbf{w}})$ for accurate UE. Refer to Appendix A.3 for proofs and further explanations. Therefore, similar to that done in Eq.(7), the intractable posterior in Eq.(4) can be approximated by $Dir(\boldsymbol{\nu}|\boldsymbol{\alpha}_{\text{ins}}; \hat{\boldsymbol{\theta}}_{\mathbf{w}})$. As a result, we can obtain a closed-form analytical solution (Appendix B.2) for instance-level uncertainty quantification.

Proposition 4.3. *Let $\mathcal{L}(\boldsymbol{\alpha}, y)$ be a loss function w.r.t $\boldsymbol{\alpha}$ and y. For any positive bag $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_K\}$, assume $\bar{w}_k \geq 0\ \forall k \in [1, K]$, $\sum_k \bar{w}_k = 1$, and $\widetilde{\boldsymbol{\alpha}} = \sum_k \bar{w}_k \boldsymbol{\alpha}_k$. $\mathcal{L}^+_{ins} = \mathcal{L}(\widetilde{\boldsymbol{\alpha}}, 1) \leq \sum_k \bar{w}_k \mathcal{L}(\boldsymbol{\alpha}_k, 1) \leq \sum_k \frac{1}{K}\mathcal{L}(\boldsymbol{\alpha}_k, 1)$ holds in instance evidential learning, when $\mathcal{L}$ is a convex function w.r.t $\boldsymbol{\alpha}$ and there is $\bar{w}_1 \geq \bar{w}_2 \geq \cdots \geq \bar{w}_K$ for $\mathcal{L}(\boldsymbol{\alpha}_1, 1) \leq \mathcal{L}(\boldsymbol{\alpha}_2, 1) \leq \cdots \leq \mathcal{L}(\boldsymbol{\alpha}_K, 1)$.*

Complete objective function To jointly train instance-level $R(\mathbf{x})$ and bag-level $S(X)$ for MIUE, the complete objective function we adopt is summarized as follows:

$$\mathcal{L} = \mathcal{L}_{\mathcal{I}\text{-EDL}} + \mathcal{L}_{\text{MIREL}} + \mathcal{L}_{\text{RED}}, \tag{11}$$

where $\mathcal{L}_{\text{RED}}$ is a RED loss (Pandey & Yu, 2023) serving as a regularization term in EDL. As $\mathcal{L}_{\text{RED}}$ shows to be effective in avoiding zero-evidence regions to improve EDL, we exploit this loss to regularize both the evidence output of $R(\mathbf{x})$ and $S(X)$. For an evidence output $\boldsymbol{\alpha}$, $\mathcal{L}_{\text{RED}} = -\frac{C}{\alpha_0}\log(\alpha_{\text{gt}} - 1)$, where α_{gt} is the predicted evidence for ground truth class. Refer to Appendix E.2 and F.3 for the ablation studies on Eq.(11). Note that in joint training, the parameters ψ and ϕ are optimized by $\mathcal{L}_{\mathcal{I}\text{-EDL}}$ in bag-level evidential learning. Meanwhile, the parameters ψ and π are optimized by $\mathcal{L}_{\text{MIREL}}$ in instance-level residual evidential learning. Thus, ψ is jointly optimized to improve both instance-level and bag-level UE performance.

5. Related Work

Uncertainty estimation (UE) methods could be grouped into two categories. (1) Bayesian NNs (**BNNs**) generally adopt stochastic NN weights to model predictive uncertainty (Neal, 2012). Many techniques (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Loquercio et al., 2020) are proposed to approximate the intractable posterior of BNNs. Nonetheless, they usually require sampling and are demanding computationally. (2) Deterministic uncertainty methods (**BUMs**) emerge as a promising means to mitigate this using deterministic NN weights. They can accomplish UE with a single forward pass mainly by regularizing hidden feature spaces (Alemi et al., 2018; Charpentier et al., 2020) or employing distance-aware output layers (Liu et al., 2020; Van Amersfoort et al., 2020; Mukhoti et al., 2023). BUM

is rapidly developing and encompasses many interesting works beyond those mentioned here. Readers could refer to Postels et al. (2022) and Mukhoti et al. (2023).

Dirichlet-based uncertainty (DBU) methods (Ulmer et al., 2023) belong to another type of BUMs. They predict the parameters of Dirichlet distribution, allowing us to distinguish different uncertainty sources and write common uncertainty measures with closed-form solutions. As one of them, EDL (Sensoy et al., 2018) have received great attention due to its simplicity and impressive performance. It has inspired many real-world applications (Bao et al., 2022; Chen et al., 2022; 2023; Zhao et al., 2023). In addition, EDL has been further improved recently in learning strategies (Deng et al., 2023; Pandey & Yu, 2023) and concepts (Fan et al., 2023), showing better results in UE. All of these motivate us to study a baseline approach for MIUE through EDL.

MIL with instance-level estimator predicts not only bag results but also instance responses. It contains two main classes. (1) Instance-level methods (Liu et al., 2012). For model interpretability, they directly predict instance scores and then aggregate these scores into a bag-level result (Ilse et al., 2020). (2) Embedding-level methods with explicit instance scoring branches. They work on instance embeddings rather than scores, and aim to enhance the accuracy of both instance and bag prediction using two separate branches (Chikontwe et al., 2020; Qu et al., 2022) or two branches decoupled via attention (Ilse et al., 2018; Shi et al., 2020; Li et al., 2021; Cui et al., 2023). Most of them are specially designed for pathology applications. In addition, some other methods concentrate on alleviating the negative effect of noisy instances on MIL to improve the accuracy, mainly by maximizing the gap between two representative instances from a pair of negative and positive bags (Tian et al., 2021; Sapkota & Yu, 2022). These methods are often seen in video anomaly detection tasks. This paper roughly follows the second class; yet we do not focus on classification accuracy but UE performance, and devise a new residual instance estimator independent of attention-based MIL.

Uncertainty estimation for MIL is formally studied far less than that for standard single instance learning, largely due to the weak-supervision nature of MIL. A recent orthogonal work is based on BNNs (Schmidt et al., 2023; Cui et al., 2023). It converts ABMIL networks into BNNs and model the uncertainty of attention scores by variational approximation. However, it focuses on improving classification performance, not specially for UE. Moreover, it also requires multiple forward passes like typical BNNs. By contrast, we propose a DBU method that quantifies the uncertainty in MIL with a single forward pass. In particular, our study focuses on UE and designs extensive experiments to assess the UE capability of MIL models.

6. Experiments

6.1. Experimental Setup

Datasets (1) Two bag datasets are **MNIST-bags** (LeCun, 1998) and **CIFAR10-bags** (Krizhevsky et al., 2009), following Ilse et al. (2018) to generate bags for MIL. A bag is positive if it contains at least one instance with the class of interest. The class of interest is '9' for MNIST and 'truck' for CIFAR10. FMNIST-bags (Xiao et al., 2017), KMNIST-bags (Clanuwat et al., 2018), SVHN-bags (Yuval, 2011), and Texture-bags (Cimpoi et al., 2014) are taken as OOD (out-of-distribution) datasets. (2) One pathology dataset is **CAMELYON16** (Bejnordi et al., 2017) for breast cancer metastasis detection. We synthesize its three distribution-shifted versions (Tellez et al., 2019) for detection and take TCGA-PRAD (Kandoth et al., 2013) as OOD dataset. More details are provided in Appendix D.1 and D.2.

Baselines (1) **Classical deep MIL networks**: Mean, Max, ABMIL (Ilse et al., 2018), and DSMIL (Li et al., 2021). They cover three popular MIL pooling operators, used to verify whether our MIREL could improve their UE performances. (2) **Related UE methods**. General ones, Deep Ensemble (Lakshminarayanan et al., 2017), MC Dropout (Gal & Ghahramani, 2016), and $\mathcal{I}$-EDL (Deng et al., 2023) are adopted. ABMIL is employed as the base network for them and our MIREL. Moreover, a BNN-based MIL method, Bayes-MIL (Cui et al., 2023), is also compared. Refer to Appendix D.3 for implementation and training details.

Evaluation Both bag-level and instance-level uncertainty are quantified for evaluation. We mainly use two typical UE tasks, confidence evaluation (**Conf.**) and **OOD detection**, *i.e.*, we assess whether a model could show more confidence (or less uncertainty) for those correctly-classified samples (*vs.* misclassified ones) and those ID (in-distribution) samples (*vs.* OOD ones). AUROC is their evaluation metrics. We calculate max probability as confidence measure by default. Following Deng et al. (2023), for EDL models, we adopt Max.α ($\max_c \alpha_c$) and α_0 ($\sum_c \alpha_c$) as confidence measures in Conf. and OOD detection, respectively. Classification accuracy (**Acc.**) is listed *only for reference*. Each model is run with 5 seeds, and we report the mean and standard deviation of evaluation metrics.

6.2. MNIST-bags

Main results are shown in Table 1. (1) **Comparing the deep MIL networks with and without our MIREL**, we have three main observations. (i) In instance-level Conf., our MIREL always helps existing MIL networks to perform better, with an average improvement of 16.49%. (ii) In bag-level OOD detection, the MIL networks with our MIREL exceed their counterparts by 7.03% on average, in 7 out of 8 comparisons. At instance level, they present an average

Table 1. Main results on **MNIST-bags**. **OOD-F** and **OOD-K** mean that FMNIST and KMNIST are used for generating OOD bags, respectively. The results colored in gray are from our derived instance estimator $T(\mathbf{x})$. $\overline{\boldsymbol{UE}}$ is the average metrics on three UE tasks.

Method	Bag-level					Instance-level				
	Acc.	Conf.	OOD-F	OOD-K	$\overline{UE}$	Acc.	Conf.	OOD-F	OOD-K	$\overline{UE}$
- Combined with deep MIL networks										
Mean	93.38 ± 0.90	**87.02** ± 1.04	**77.57** ± 2.46	54.66 ± 2.62	73.08	86.52 ± 0.97	66.49 ± 1.37	**79.36** ± 1.95	**57.43** ± 1.50	67.76
Mean + MIREL	93.50 ± 0.53	87.01 ± 1.04	75.26 ± 1.52	**57.69** ± 6.28	**73.32**	92.45 ± 1.22	**91.49** ± 1.76	69.98 ± 4.41	56.70 ± 4.97	**72.72**
Max	94.56 ± 0.46	87.82 ± 1.49	75.23 ± 1.32	62.44 ± 3.00	75.17	92.53 ± 0.54	81.86 ± 1.54	76.97 ± 1.71	**62.53** ± 1.61	73.79
Max + MIREL	95.96 ± 0.29	**87.85** ± 2.23	**84.17** ± 3.32	**66.75** ± 5.70	**79.59**	96.82 ± 0.27	**84.22** ± 0.43	**80.81** ± 4.88	61.15 ± 3.28	**75.40**
DSMIL	96.22 ± 0.17	**87.56** ± 0.95	71.13 ± 5.20	60.71 ± 7.91	73.13	70.16 ± 3.56	64.64 ± 0.49	59.75 ± 2.35	57.50 ± 2.55	60.63
DSMIL + MIREL	96.50 ± 0.37	87.26 ± 2.66	**87.27** ± 4.27	**62.03** ± 7.78	**78.85**	97.19 ± 0.29	**73.79** ± 15.68	**73.29** ± 10.85	**57.58** ± 3.44	**68.22**
ABMIL	95.74 ± 0.38	**86.91** ± 0.98	82.93 ± 4.81	74.37 ± 4.84	81.41	75.03 ± 0.28	61.28 ± 0.86	63.68 ± 1.00	52.63 ± 1.07	59.20
ABMIL + MIREL	96.48 ± 0.22	86.63 ± 1.32	**92.84** ± 0.60	**79.95** ± 4.12	**86.47**	87.71 ± 0.67	**90.73** ± 1.31	**78.13** ± 2.19	**67.02** ± 1.94	**78.63**
- Compared with related UE methods using ABMIL as the base MIL network										
Deep Ensemble	96.06 ± 0.35	87.36 ± 0.59	80.07 ± 2.57	74.33 ± 3.97	80.59	75.56 ± 0.32	71.89 ± 0.91	70.48 ± 0.53	55.22 ± 1.16	65.87
MC Dropout	96.28 ± 0.41	**88.46** ± 1.82	89.57 ± 3.84	78.24 ± 4.89	85.42	75.61 ± 0.66	68.40 ± 1.54	68.34 ± 1.06	58.61 ± 1.38	65.12
$\mathcal{I}$-EDL	96.08 ± 0.20	86.78 ± 0.87	85.51 ± 7.56	73.15 ± 3.87	81.82	75.45 ± 0.13	60.72 ± 1.46	63.91 ± 1.31	54.14 ± 2.19	59.59
Bayes-MIL	96.44 ± 0.33	85.63 ± 1.53	81.02 ± 11.71	57.04 ± 12.61	74.57	91.64 ± 1.25	82.24 ± 1.85	60.77 ± 6.59	42.06 ± 2.84	61.69
MIREL	96.48 ± 0.22	86.63 ± 1.32	**92.84** ± 0.60	**79.95** ± 4.12	**86.47**	87.71 ± 0.67	**90.73** ± 1.31	**78.13** ± 2.19	**67.02** ± 1.94	**78.63**

improvement of 9.26%, in 5 out of 8 comparisons. (iii) Overall, our MIREL could often enhance the performance of deep MIL networks by large margins in terms of UE, especially for Max, ABMIL, and DSMIL. (2) **Compared with related UE methods**, our MIREL always shows better UE performances except in bag-level Conf.; particularly, at instance level, our MIREL leads runner-up by 7.62% $\sim$ 8.41%. These comparative results suggest that our MIREL is an effective and preferable approach for MIUE.

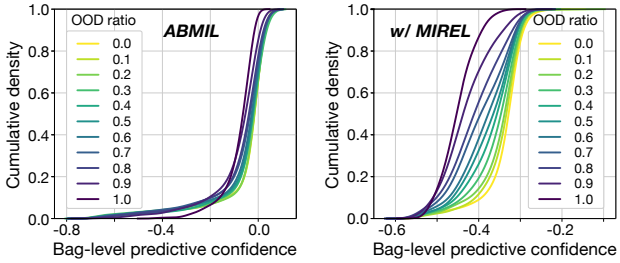

Figure 1. Distribution of bag-level predictive confidence (negative expected entropy). Different ratios of OOD (FMNIST) instances are set to assess the UE capability of MIL models.

Uncertainty analysis is carried out to further examine the UE capability of our method. (1) **Bag-level**. We use OOD instances to randomly replace the instances from ID test bags, according to a specific target ratio of OOD instances. From the results shown in Fig. 1, we find that it is hard for ABMIL to identify the bags with different degrees of anomalies; but interestingly, when combined with our MIREL, ABMIL shows plausible reflections on those various abnormal bags. (2) **Instance-level**. We show results in Fig. 2. (i) From the result at top row, we observe that the ABMIL with our MIREL tends to predict clearly-higher uncertainty for numbers '4' and '7' while ABMIL does so only for '7'. The former result is in line with experiences since both '4' and '7' can be easily mistaken with '9' in hand-written numbers (Ilse et al., 2018). (ii) From the result at bottom row, we see that the ABMIL w/ MIREL is more likely to predict lower confidences for OOD instances than ABMIL. These results further confirm that our MIREL could assist classical MIL networks to capture uncertainty. More uncertainty analysis on KMNIST, α_0, and DSMIL are given in Appendix E.1.

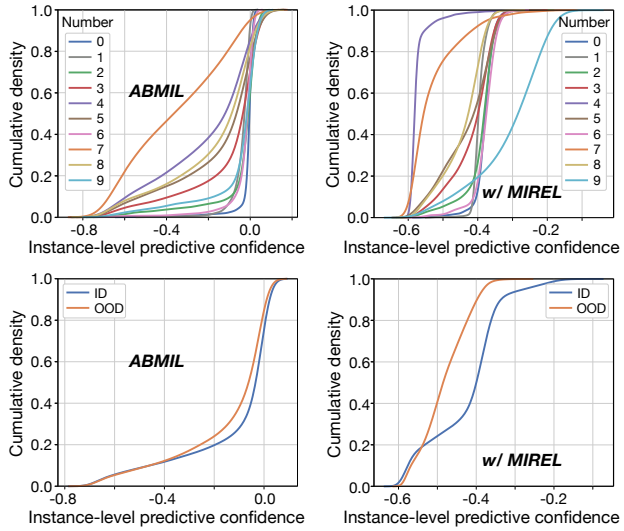

Figure 2. Distribution of instance-level predictive confidence (negative expected entropy). Top row is the result of MNIST instances, where '9' is the number of interest (positive instance). Bottom row is the result of ID (MNIST) and OOD (FMNIST) instances.

Ablation study is conducted on our MIREL to verify the

Table 2. Ablation study on the ABMIL with our MIREL using **MNIST-bags**.

Loss		Ins.	Bag-level					Instance-level				
$\mathcal{L}_{\mathcal{I}\text{-EDL}}$	$\mathcal{L}_{\text{MIREL}}$		Acc.	Conf.	OOD-F	OOD-K	$\overline{UE}$	Acc.	Conf.	OOD-F	OOD-K	$\overline{UE}$
		a_k	95.74 $\pm$ 0.38	**86.91** $\pm$ 0.98	82.93 $\pm$ 4.81	74.37 $\pm$ 4.84	81.41	75.03 $\pm$ 0.28	61.28 $\pm$ 0.86	63.68 $\pm$ 1.00	52.63 $\pm$ 1.07	59.20
		T	95.74 $\pm$ 0.38	**86.91** $\pm$ 0.98	82.93 $\pm$ 4.81	74.37 $\pm$ 4.84	81.41	87.03 $\pm$ 1.42	84.30 $\pm$ 3.45	64.64 $\pm$ 4.92	52.67 $\pm$ 4.23	67.20
✓		a_k	96.08 $\pm$ 0.20	86.78 $\pm$ 0.87	85.51 $\pm$ 7.56	73.15 $\pm$ 3.87	81.82	75.45 $\pm$ 0.13	60.72 $\pm$ 1.46	63.91 $\pm$ 1.31	54.14 $\pm$ 2.19	59.59
✓		T	96.08 $\pm$ 0.20	86.78 $\pm$ 0.87	85.51 $\pm$ 7.56	73.15 $\pm$ 3.87	81.82	85.19 $\pm$ 0.64	87.67 $\pm$ 1.11	73.52 $\pm$ 5.66	56.63 $\pm$ 1.66	72.60
✓	✓	R	96.48 $\pm$ 0.22	86.63 $\pm$ 1.32	**92.84** $\pm$ 0.60	**79.95** $\pm$ 4.12	**86.47**	87.71 $\pm$ 0.67	**90.73** $\pm$ 1.31	**78.13** $\pm$ 2.19	**67.02** $\pm$ 1.94	**78.63**

Table 3. Comparison with related UE methods on **CAMELYON16**. **OOD-PRAD** means that TCGA-PRAD is taken as OOD data. The baseline of this experiment is vanilla ABMIL without any additional UE techniques. $\overline{UE}$ is the average metrics on two UE tasks.

Method	Bag-level				Instance-level			
	Acc.	Conf.	OOD-PRAD	$\overline{UE}$	Acc.	Conf.	OOD-PRAD	$\overline{UE}$
Baseline	86.77 $\pm$ 0.77	**72.05** $\pm$ 1.72	41.90 $\pm$ 2.91	56.98	96.07 $\pm$ 0.01	49.87 $\pm$ 3.28	31.34 $\pm$ 0.85	40.60
Deep Ensemble	86.93 $\pm$ 0.63	70.44 $\pm$ 1.79	39.62 $\pm$ 3.32	55.03	96.08 $\pm$ 0.02	49.62 $\pm$ 2.53	28.16 $\pm$ 1.07	38.89
MC Dropout	87.09 $\pm$ 1.37	67.50 $\pm$ 5.92	45.66 $\pm$ 5.48	56.58	96.05 $\pm$ 0.00	56.35 $\pm$ 2.20	33.93 $\pm$ 2.05	45.14
$\mathcal{I}$-EDL	87.72 $\pm$ 0.63	57.48 $\pm$ 8.07	72.43 $\pm$ 11.98	64.95	96.05 $\pm$ 0.01	45.41 $\pm$ 5.33	32.06 $\pm$ 1.49	38.74
Bayes-MIL	86.61 $\pm$ 1.11	66.91 $\pm$ 6.73	48.77 $\pm$ 7.97	57.84	97.27 $\pm$ 0.28	**82.12** $\pm$ 9.12	54.53 $\pm$ 21.34	68.33
MIREL	87.09 $\pm$ 1.07	61.62 $\pm$ 6.85	**82.51** $\pm$ 8.34	**72.06**	97.79 $\pm$ 0.71	77.85 $\pm$ 5.69	**67.85** $\pm$ 5.71	**72.85**

effectiveness of its components. Different loss functions and instance estimators are adopted. More details are shown in Appendix D.3. From the results shown in Table 2, there are three main findings. (1) **For our derived instance estimator** T, it leads attention-based scoring proxy (a_k) at instance level by large margins (8% and 13.01%) in terms of overall UE performance. (2) **For the adopted** $\mathcal{L}_{\mathcal{I}\text{-EDL}}$, it could help to enhance instance-level UE performances in the presence of T; no obvious effect is observed in other cases. (3) **For our new residual estimator** R trained with $\mathcal{L}_{\text{MIREL}}$, it boosts not only instance-level but also bag-level UE performances. Particularly, its performance improvements in OOD detection range from 4.61% to 10.39%. These findings could demonstrate the effectiveness of those components proposed in our MIREL. **More studies**, including i) the optimization strategies for $R(\mathbf{x})$ and ii) adopting $T(\mathbf{x})$ for related UE methods, are presented in Appendix E.2.

More experiments (1) The results on **CIFAR10-bags**, including main results (F.1), uncertainty analysis (F.2), and ablation studies (F.3), are given in Appendix F. **(2) A synthetic MIUE experiment** is presented in Appendix H, in order to intuitively understand the behavior of our weakly-supervised residual instance estimator $R(\mathbf{x})$.

6.3. Histopathology Dataset

Main results are exhibited in Table 3. We mainly compare our method with related UE methods in this experiment. There are four main observations from these results. (1) Our MIREL obtains the best overall UE performance at both bag and instance levels, surpassing the second best method by 7.11% and 4.52%, respectively. (2) Our MIREL improves the overall performance of ABMIL (baseline) in UE by considerable margins, 15.08% and 32.35% at bag and instance level, respectively. (3) Especially, our method shows impressive results in OOD detection, with a margin of more than 10% over others at both instance and bag level. (4) The UE methods not specially proposed for MIL often obtain the AUROC even less than 0.5 in OOD detection. It is because, the pathology images from TCGA-PRAD are near-OOD data and near-OOD detection is usually more challenging than far-OOD detection for these UE methods. These four observations suggest that our MIREL has the potential to be applied in real-world applications.

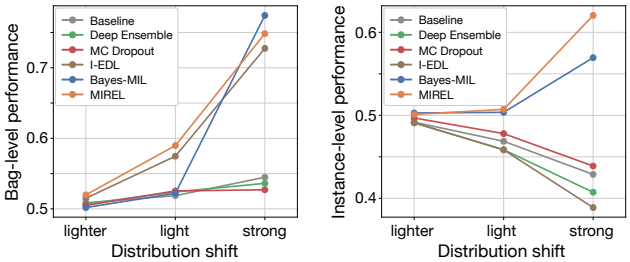

Figure 3. Performance of compared UE methods in distribution shift detection on **CAMELYON16**. The test samples of CAMELYON16 are shifted with different degrees of image noises for detection. Refer to Appendix G.2 for complete numerical results.

Distribution shift (DS) detection is a more challenging task than OOD detection, so it is further adopted to test our method. We generate three distribution-shifted test sets by adding routine noises to original CAMELYON16 test images, called *lighter*, *light*, and *strong* according to noise strengths. Refer to Appendix D.2 for experimental details. Detection result is shown in Fig. 3. (1) **Bag-level**. Lighter DS is hard to detect for all the adopted UE methods (only

about 0.52). Our MIREL can detect both light and strong DS with the best or the second best performance, while the competitive Bayes-MIL fails to detect light DS with an AUROC less than 0.53. (2) **instance-level**. All the compared methods *not specially* for MIL, consistently show meaningless results (AUROC < 0.5). By contrast, our MIREL detects strong DS with an AUROC of 0.62, better than the Bayes-MIL with an AUROC of 0.57. This experiment could further verify the superiority of our MIREL in MIUE.

7. Limitation and Future Work

Although our baseline approach MIREL shows promising results in MIUE, there are still some limitations in our experiments. First, since multi-instance bag is usually expensive in computation, our MIL datasets are limited in scale. Larger datasets (*e.g.*, with more than 10,000 bags) would be better for more comprehensive validation. In addition, we take AUROC as the main metric to quantitatively evaluate the performance of UE methods. Additional calibration metrics, *e.g.*, Expected Calibration Error (ECE) or Brier Score, would help to evaluate more holistically.

In the future, there are some directions worth further research. (1) The optimization strategy for the weakly-supervised posterior $p(\boldsymbol{\theta}_{\mathbf{w}}|\mathcal{D})$. It could be further improved to provide a tighter upper bound for more accurate UE at instance level. (2) Seeking other efficient UE methods, such as distance-aware UE and feature space regularization, as stated in Postels et al. (2022). (3) More general settings beyond binary MIL, *e.g.*, multi-label MIL, as they cover more practical applications (Zhou et al., 2012).

8. Conclusion

This paper addresses a new MIUE problem and presents a baseline scheme, *Multi-Instance Residual Evidential Learning*. In this scheme, we propose to model bag-level predictive uncertainty using a Dirichlet-based posterior distribution parameterized by general MIL networks. In particular, at weakly-supervised instance level, we derive a new residual estimator through the Fundamental Theorem of Symmetric Functions for instance-level UE. Moreover, without complete instance labels, we propose a weakly-supervised evidential optimization strategy for that residual estimator. Different UE tasks and extensive experiments demonstrate that our MIREL could often outperform other related UE methods. In addition, it can be applied to existing MIL networks, effectively assisting them in improving MIUE performances, especially at instance level. Since MIL has close connections with many real-world and safe-critical applications, it is of great importance and highly anticipated to enhance the reliability of MIL systems through MIUE. Our work could inspire more research on investigating MIUE, paving the way to explore uncertainty estimation in more weakly-supervised settings.

Acknowledgements

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant ZYGX2022YGRH015 and in part by the National Natural Science Foundation of China (NSFC) under Grant 61972072. The authors would like to thank Jianghua Tian and Mao Dai for their encouragement and helpful feedback, and anonymous reviewers for their constructive comments to improve this work.

Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alemi, A. A., Fischer, I., and Dillon, J. V. Uncertainty in the variational information bottleneck. *arXiv preprint arXiv:1807.00906*, 2018.

Babenko, B., Yang, M.-H., and Belongie, S. Robust object tracking with online multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1619–1632, 2010.

Bao, W., Yu, Q., and Kong, Y. Opental: Towards open set temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2979–2989, 2022.

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J. A., Hermsen, M., Manson, Q. F., Balkenhol, M., et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

Blanchard, G., Lee, G., and Scott, C. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.

Charpentier, B., Zügner, D., and Günnemann, S. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in Neural Information Processing Systems*, 33:1356–1367, 2020.

Chen, M., Gao, J., Yang, S., and Xu, C. Dual-evidential learning for weakly-supervised temporal action localiza-

tion. In *European Conference on Computer Vision*, pp. 192–208. Springer, 2022.

Chen, M., Gao, J., and Xu, C. Cascade evidential learning for open-world weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14741–14750, 2023.

Chikontwe, P., Kim, M., Nam, S. J., Go, H., and Park, S. H. Multiple instance learning with center embeddings for histopathology classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pp. 519–528. Springer, 2020.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Cui, Y., Liu, Z., Liu, X., Liu, X., Wang, C., Kuo, T.-W., Xue, C. J., and Chan, A. B. Bayes-MIL: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In *The Eleventh International Conference on Learning Representations*, 2023.

Dempster, A. P. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.

Deng, D., Chen, G., Yu, Y., Liu, F., and Heng, P.-A. Uncertainty estimation by Fisher information-based evidential deep learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 7596–7616. PMLR, 2023.

Der Kiureghian, A. and Ditlevsen, O. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

Fan, L., Liu, B., Li, H., Wu, Y., and Hua, G. Flexible visual recognition by evidential modeling of confusion and ignorance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1338–1347, 2023.

Franchi, G., Yu, X., Bursuc, A., Tena, A., Kazmierczak, R., Dubuisson, S., Aldea, E., and Filliat, D. Muad: Multiple uncertainties for autonomous driving, a benchmark for multiple uncertainty types and tasks. *arXiv preprint arXiv:2203.01437*, 2022.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Ilse, M., Tomczak, J., and Welling, M. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

Ilse, M., Tomczak, J. M., and Welling, M. Deep multiple instance learning for digital histopathology. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, pp. 521–546. Elsevier, 2020.

Jøsang, A. *Subjective logic*, volume 3. Springer, 2016.

Kandemir, M. and Hamprecht, F. A. Computer-aided diagnosis from weak supervision: A benchmarking study. *Computerized medical imaging and graphics*, 42:44–50, 2015.

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., et al. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471): 333–339, 2013.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Li, B., Li, Y., and Eliceiri, K. W. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14318–14328, 2021.

Li, H., Zhu, C., Zhang, Y., Sun, Y., Shui, Z., Kuang, W., Zheng, S., and Yang, L. Task-specific fine-tuning via variational information bottleneck for weakly-supervised

pathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7454–7463, 2023.

Linmans, J., Elfwing, S., van der Laak, J., and Litjens, G. Predictive uncertainty estimation for out-of-distribution detection in digital pathology. *Medical Image Analysis*, 83:102655, 2023.

Liu, G., Wu, J., and Zhou, Z.-H. Key instance detection in multi-instance learning. In *Asian conference on machine learning*, pp. 253–268. PMLR, 2012.

Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., and Lakshminarayanan, B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020.

Liu, P., Ji, L., Ye, F., and Fu, B. AdvMIL: Adversarial multiple instance learning for the survival analysis on whole-slide images. *Medical Image Analysis*, 91:103020, 2024a.

Liu, P., Ji, L., Zhang, X., and Ye, F. Pseudo-bag mixup augmentation for multiple instance learning-based whole slide image classification. *IEEE Transactions on Medical Imaging*, 43(5):1841–1852, 2024b.

Loquercio, A., Segu, M., and Scaramuzza, D. A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters*, 5(2):3153–3160, 2020.

Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

Mena, J., Pujol, O., and Vitria, J. A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective. *ACM Computing Surveys (CSUR)*, 54(9):1–35, 2021.

Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H., and Gal, Y. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24384–24394, 2023.

Neal, R. M. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–436, 2015. doi: 10.1109/CVPR.2015.7298640.

Pandey, D. S. and Yu, Q. Learn to accumulate evidence from all training samples: theory and practice. In *International Conference on Machine Learning*, pp. 26963–26989. PMLR, 2023.

Postels, J., Segù, M., Sun, T., Sieber, L. D., Van Gool, L., Yu, F., and Tombari, F. On the practicality of deterministic epistemic uncertainty. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17870–17909. PMLR, 2022.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

Qu, L., Wang, M., Song, Z., et al. Bi-directional weakly supervised knowledge distillation for whole slide image classification. *Advances in Neural Information Processing Systems*, 35:15368–15381, 2022.

Rizve, M. N., Mittal, G., Yu, Y., Hall, M., Sajeev, S., Shah, M., and Chen, M. Pivotal: Prior-driven supervision for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22992–23002, June 2023.

Sapkota, H. and Yu, Q. Bayesian nonparametric submodular video partition for robust anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3212–3221, 2022.

Schmidt, A., Morales-Álvarez, P., and Molina, R. Probabilistic attention based on gaussian processes for deep multiple instance learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Shi, X., Xing, F., Xie, Y., Zhang, Z., Cui, L., and Yang, L. Loss-based attention for deep multiple instance learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5742–5749, 2020.

Sultani, W., Chen, C., and Shah, M. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488, 2018.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., and Van Der Laak, J. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58:101544, 2019.

Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J. W., and Carneiro, G. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4975–4986, 2021.

Ulmer, D. T., Hardmeier, C., and Frellsen, J. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=xqS8k9E75c.

Van Amersfoort, J., Smith, L., Teh, Y. W., and Gal, Y. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pp. 9690–9700. PMLR, 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yuval, N. Reading digits in natural images with unsupervised feature learning. In *Proceedings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Zhao, C., Du, D., Hoogs, A., and Funk, C. Open set action recognition via multi-label evidential learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22982–22991, 2023.

Zhong, J.-X., Li, N., Kong, W., Liu, S., Li, T. H., and Li, G. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1237–1246, 2019.

Zhou, Z.-H., Zhang, M.-L., Huang, S.-J., and Li, Y.-F. Multi-instance multi-label learning. *Artificial Intelligence*, 176 (1):2291–2320, 2012.

A. Derivation and Proof

In this section, A.1 gives the derivation of Eq.(7) to clarify the connection of our bag-level predictive uncertainty modeling with Bayesian frameworks. A.2 presents the proof details of Corollary 4.1 and Proposition 4.2 (in Section 4.2). Lastly, in A.3, we justify our weakly-supervised optimization strategy (in Section 4.3) that aims to provide a more suitable $\hat{\boldsymbol{\theta}}_{\mathbf{w}}$.

A.1. Derivation of Eq.(7)

From the perspective of Bayesian (Neal, 2012), for a new bag input X^*, its predictive distribution can be written by Eq.(3):

$$P(Y^*|X^*, \mathcal{D}) = \int P(Y^*|X^*, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\mathcal{D}) d\boldsymbol{\omega}.$$

By introducing a new distribution $p(\boldsymbol{\mu}|X^*, \boldsymbol{\omega})$, we can re-write the equation above as follows:

$$\begin{aligned} \int P(Y^*|X^*, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\mathcal{D}) d\boldsymbol{\omega} &= \int \int P(Y^*|\boldsymbol{\mu}) p(\boldsymbol{\mu}|X^*, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\mathcal{D}) d\boldsymbol{\mu} d\boldsymbol{\omega} \\ &= \int P(Y^*|\boldsymbol{\mu}) \Big[\int p(\boldsymbol{\mu}|X^*, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\mathcal{D}) d\boldsymbol{\omega} \Big] d\boldsymbol{\mu} \\ &= \int P(Y^*|\boldsymbol{\mu}) p(\boldsymbol{\mu}|X^*, \mathcal{D}) d\boldsymbol{\mu} \end{aligned} \tag{12}$$

As a result, $P(Y^*|\boldsymbol{\mu})$ can be taken as the new model. $p(\boldsymbol{\mu}|X^*, \mathcal{D})$ is the distribution over model parameters conditioned the input bag X and the given bag dataset $\mathcal{D}$. It is also referred to as the estimate of *distributional uncertainty* given model uncertainty. However, the marginalization of the equation above is intractable. To tackle this, a point estimate of model parameters $\hat{\boldsymbol{\omega}}$ is often assumed to satisfy $p(\boldsymbol{\omega}|\mathcal{D}) = \delta(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}})$ (Malinin & Gales, 2018). Hence,

$$\int P(Y^*|X^*, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\mathcal{D}) d\boldsymbol{\omega} = \int P(Y^*|\boldsymbol{\mu}) p(\boldsymbol{\mu}|X^*, \mathcal{D}) d\boldsymbol{\mu} \approx \int P(Y^*|\boldsymbol{\mu}) p(\boldsymbol{\mu}|X^*, \hat{\boldsymbol{\omega}}) d\boldsymbol{\mu}.$$

$p(\boldsymbol{\mu}|X^*, \hat{\boldsymbol{\omega}})$ is exactly the posterior Dirichlet distribution given in our bag-level predictive uncertainty modeling. Accordingly, quantifying its uncertainty becomes tractable, as there is a closed-form analytical solution for those commonly-used uncertainty measures (refer to Appendix B for details).

A.2. Proof of Corollary 4.1 and Proposition 4.2

***Corollary** 4.1* (to Theorem 2.1). Given a scoring function for a set of instances $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$, written as $S(X) = g\big(\sum_k f(\mathbf{x}_k)\big) \in \mathbb{R}$ where $k \in [1, K]$, a scoring function for any single instance can be written as

$$T = g \circ f,$$

and $T(\mathbf{x}) \in \mathbb{R}$ for an instance $\mathbf{x}$.

Proof. Without loss of generality, we assume $f(\mathbf{x}) \in \mathbb{R}^M$. Given that the input of $g(\cdot)$ is $\sum_k f(\mathbf{x}_k)$ in $S(X)$, the domain of $g(\cdot)$ is in $\mathbb{R}^M$, since $\sum_k f(\mathbf{x}_k) \in \mathbb{R}^M$. Therefore, $g(\cdot)$ can take $f(\mathbf{x})$ as input. Namely, there is a feasible function $T(\mathbf{x}) = g(f(\mathbf{x})) \in \mathbb{R}$. Further, $S(X)$ is stated as a bag scoring function in Theorem 2.1. Hence, the $T(\mathbf{x})$, which has the same decision function $g(\cdot)$ as $S(X)$, can also be cast a scoring function specially for instances. □

***Proposition** 4.2*. Let $S(\cdot)$ be a classifier for a bag of instances $X = \{\mathbf{x}_k\}_{k=1}^K$ and satisfy $S(X) = g\big(\sum_k f(\mathbf{x}_k)\big)$. For any bag X and its label $Y \in \{0, 1\}$, further assume S can predict bags precisely: $S(X) = Y$. Then, there exists an estimator with $T = g \circ f$ for any single instance $\mathbf{x}$, such that $T(\mathbf{x}) = y$, where $y \in \{0, 1\}$ is the label of $\mathbf{x}$.

Proof. According to Corollary 4.1, given $S(X) = g\big(\sum_k f(\mathbf{x}_k)\big)$, there is an instance-level estimator T that can be written as $T(\mathbf{x}) = g(f(\mathbf{x}))$. With the existence of T, we need to prove $T(\mathbf{x}) = y$, for any single instance $\mathbf{x}$ and its label y.

Recall that, Theorem 2.1, which gives the bag scoring function $S(X) = g\big(\sum_k f(\mathbf{x}_k)\big)$, holds as before or under weak conditions, when the form of instance pooling, $\sum_k f(\mathbf{x}_k)$, is replaced by others, such as i) mean, ii) $\max$ (Qi et al., 2017), and iii) attention-based MIL pooling (Ilse et al., 2018). Next, we finish the proof with the basics of MIL.

First of all, for any instance $\mathbf{x}$, we assume that there is a virtual bag as follows:

$$X_{\text{vir}} = \{\underbrace{\mathbf{x}, \cdots, \mathbf{x}}_{n}\}.$$

Based on the classical MIL assumption, *i.e.*, $Y = \max_k\{y_k\}$, we have

$$Y_{\text{vir}} = \max_{\mathbf{x}\in X_{\text{vir}}}\{y\} = y,$$

where Y_{vir} is the ground-truth of X_{vir}. Then, we discuss three cases for the form of instance pooling in MIL:

Case (1): mean,

$$S(X_{\text{vir}}) = g\Big(\frac{1}{n}\sum_{\mathbf{x}\in X_{\text{vir}}} f(\mathbf{x})\Big) = g(f(\mathbf{x})).$$

Case (2): $\max$ (Qi et al., 2017),

$$S(X_{\text{vir}}) = g\Big(\max_{\mathbf{x}\in X_{\text{vir}}}\{f(\mathbf{x})\}\Big) = g(f(\mathbf{x})).$$

Case (3): attention (Ilse et al., 2018). Combining with Eq.(2), there is

$$S(X_{\text{vir}}) = g\Big(\sum_{\mathbf{x}\in X_{\text{vir}}} \frac{\exp\big(t(f(\mathbf{x}))\big)}{\sum_{\tau=1}^{n}\exp\big(t(f(\mathbf{x}))\big)} f(\mathbf{x})\Big) = g\Big(\sum_{\mathbf{x}\in X_{\text{vir}}} \frac{1}{n} f(\mathbf{x})\Big) = g(f(\mathbf{x})).$$

Since there is $S(X) = Y$, we have

$$\hat{Y}_{\text{vir}} = S(X_{\text{vir}}) = Y_{\text{vir}},$$

where $\hat{Y}_{\text{vir}}$ is the prediction of X_{vir}. Eventually,

$$\left.\begin{aligned} Y_{\text{vir}} &= \max_{\mathbf{x}\in X_{\text{vir}}}\{y\} = y \\ S(X_{\text{vir}}) &= g(f(\mathbf{x})) \\ \hat{Y}_{\text{vir}} &= S(X_{\text{vir}}) = Y_{\text{vir}} \end{aligned}\right\} \implies \hat{y} = T(\mathbf{x}) = g(f(\mathbf{x})) = y \tag{13}$$

□

A.3. Justification for the Objective Function $\mathcal{L}_{\text{MIREL}}$ Given in Eq.(10)

This section presents the details of our justification for $\mathcal{L}_{\text{MIREL}}$, including the proof of Proposition 4.3.

Considering any bag $X \in \mathcal{D}$ and its label $Y \in \{0, 1\}$, there are $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_K\}$ and $Y = \max\{y_1, \cdots, y_K\}$ in MIL. To train an unbiased instance estimator $R(\mathbf{x})$ given instance labels, an ideal loss function could be written as follows:

$$\min \mathcal{L}_R = \min \sum_{k=1}^{K} \frac{1}{K}\mathcal{L}(\boldsymbol{\alpha}_k, y_k), \tag{14}$$

where $\mathcal{L}$ is a loss function derived from MLE for evidence learning, and $\boldsymbol{\alpha}_k$ is the concentration parameter of predictive Dirichlet distribution for the k-th instance. Note that here $\mathcal{L}$ is a loss function for $\boldsymbol{\alpha}_k$, *i.e.*, the prediction of the k-th instance, rather than a raw instance input.

When $Y = 0$, we have $y_k = 0\ \forall k = [1, K]$. In this case, $\mathcal{L}_R$ can be directly used for optimization given instance labels. When $Y = 1$, we only know $\max_k\{y_k\} = 1$, without complete instance labels. In this case, there are three alternative strategies for the training of $R(\mathbf{x})$.

Strategy (1): naive instance label assignment. Assuming $y_k = 1\ \forall k = [1, K]$, an alternative objective function is

$$\min \mathcal{L}_1 = \min \sum_{k=1}^{K} \frac{1}{K}\mathcal{L}(\boldsymbol{\alpha}_k, 1).$$

Strategy (2): naive instance label assignment and weighted loss. It is a weighted variant of strategy (1), written as follows:

$$\min \mathcal{L}_2 = \min \sum_{k=1}^{K} \frac{w_k}{\sum_{\tau=1}^{K} w_\tau} \mathcal{L}(\boldsymbol{\alpha}_k, 1),$$

where w_k is the probability of the k-th instance being positive. Here, w_k could be estimated by $T(\mathbf{x}_k) = g(f(\mathbf{x}_k))$.

Strategy (3): weighted instance evidence. w_k is used to aggregate instance evidences in order to learn from key positive instances. This is the strategy we adopt, as written in Eq.(10). We simplify its objective function and re-write it by

$$\min \mathcal{L}_{\text{ins}}^{+} = \min \mathcal{L}\Big(\sum_{k=1}^{K} \frac{w_k}{\sum_{\tau=1}^{K} w_\tau} \boldsymbol{\alpha}_k, 1 \Big).$$

Next, we prove that strategy (3) can provide a tighter upper bound for ideal Eq.(14) than the other two under given conditions. Before that, we first give Proposition A.1 and Proposition A.2, as well as their proof, as follows:

Proposition A.1. *$\mathcal{L}_{ins}^{+} \leq \mathcal{L}_2$ holds in instance evidential learning for a convex objective function $\mathcal{L}(\boldsymbol{\alpha}, y = 1)$.*

Proof. By definition, $w_k \geq 0$. $\mathcal{L}(\boldsymbol{\alpha}, y = 1)$ is a given convex function *w.r.t* $\boldsymbol{\alpha}$. Therefore, by Jensen's inequality we have

$$\mathcal{L}_{\text{ins}}^{+} = \mathcal{L}\Big(\sum_{k=1}^{K} \frac{w_k}{\sum_{\tau=1}^{K} w_\tau} \boldsymbol{\alpha}_k, 1 \Big) \leq \sum_{k=1}^{K} \frac{w_k}{\sum_{\tau=1}^{K} w_\tau} \mathcal{L}(\boldsymbol{\alpha}_k, 1) = \mathcal{L}_2. \tag{15}$$

□

✍ **Additional explanation:** The condition given in Proposition A.1, a convex $\mathcal{L}$ *w.r.t* model prediction, could be satisfied for common loss functions, *e.g.*, the negative logarithm *w.r.t* prediction, or the mean square error between prediction and target. Although the $\mathcal{L}(\boldsymbol{\alpha}, y = 1)$ used for our instance evidential learning (Appendix C) contains non-convex terms and is not a strict convex function *w.r.t* $\boldsymbol{\alpha}$, we still find that $\mathcal{L}_{\text{ins}}^{+}$ could be better than $\mathcal{L}_2$ in terms of overall UE performance, demonstrated by the ablation study on optimization strategy (refer to Appendix E.2 and Appendix F.3).

Proposition A.2. *$\mathcal{L}_2 \leq \mathcal{L}_1$ holds in instance evidential learning when there is $w_1 \geq w_2 \geq \cdots \geq w_K$ for $\mathcal{L}(\boldsymbol{\alpha}_1, 1) \leq \mathcal{L}(\boldsymbol{\alpha}_2, 1) \leq \cdots \leq \mathcal{L}(\boldsymbol{\alpha}_K, 1)$.*

Proof. Let

$$b_k = \frac{w_k}{\sum_{\tau=1}^{K} w_\tau} \; \forall k = 1, \cdots, K.$$

Hence $\sum_{k=1}^{K} b_k = 1$. Given $w_k \geq 0$ and $w_1 \geq w_2 \geq \cdots \geq w_K$, there is

$$b_1 \geq b_2 \geq \cdots \geq b_K.$$

First of all, we prove the following inequality through proof by contradiction:

$$\Delta_n = \sum_{k=1}^{n} b_k - \frac{n}{K} \geq 0.$$

Concretely, we assume $\Delta_n < 0$, *i.e.*, $\sum_{k=1}^{n} b_k < \frac{n}{K}$. Given $b_1 \geq b_2 \geq \cdots \geq b_n \geq b_q \; \forall q = n+1, \cdots, K$, we have

$$\sum_{k=1}^{n} b_k < \frac{n}{K} \Longrightarrow \sum_{k=1}^{n} b_q \leq \sum_{k=1}^{n} b_k < \frac{n}{K} \Longrightarrow n b_q < \frac{n}{K} \Longrightarrow b_q < \frac{1}{K}.$$

Further, by adding all b_q into $\sum_{k=1}^{n} b_k$ and using $b_q < \frac{1}{K}$, there is

$$\sum_{k=1}^{n} b_k + \sum_{q=n+1}^{K} b_q < \frac{n}{K} + \sum_{q=n+1}^{K} \frac{1}{K} \Longrightarrow \sum_{k=1}^{n} b_k + \sum_{q=n+1}^{K} b_q < \frac{n}{K} + \frac{K-n}{K} \Longrightarrow \sum_{k=1}^{n} b_k + \sum_{q=n+1}^{K} b_q < 1.$$

This contradicts $\sum_{k=1}^{K} b_k = 1$. Namely, $\Delta_n = \sum_{k=1}^{n} b_k - \frac{n}{K} \geq 0$ holds.

Then, given $\mathcal{L}(\boldsymbol{\alpha}_1, 1) \leq \mathcal{L}(\boldsymbol{\alpha}_2, 1)$ and $\Delta_1 = b_1 - \frac{1}{K} \geq 0$, there is

$$\Big(b_1 - \frac{1}{K}\Big)\mathcal{L}(\boldsymbol{\alpha}_1, 1) \leq \Big(b_1 - \frac{1}{K}\Big)\mathcal{L}(\boldsymbol{\alpha}_2, 1) \implies b_1\mathcal{L}(\boldsymbol{\alpha}_1, 1) + \Big(\frac{2}{K} - b_1\Big)\mathcal{L}(\boldsymbol{\alpha}_2, 1) \leq \frac{1}{K}\mathcal{L}(\boldsymbol{\alpha}_1, 1) + \frac{1}{K}\mathcal{L}(\boldsymbol{\alpha}_2, 1).$$

Further, by introducing $\Delta_2\mathcal{L}(\boldsymbol{\alpha}_2, 1) \leq \Delta_2\mathcal{L}(\boldsymbol{\alpha}_3, 1)$ ($\Delta_2 = b_1 + b_2 - \frac{2}{K} \geq 0$) into the inequality above, we have

$$b_1\mathcal{L}(\boldsymbol{\alpha}_1, 1) + b_2\mathcal{L}(\boldsymbol{\alpha}_2, 1) + \Big(\frac{3}{K} - b_1 - b_2\Big)\mathcal{L}(\boldsymbol{\alpha}_3, 1) \leq \frac{1}{K}\mathcal{L}(\boldsymbol{\alpha}_1, 1) + \frac{1}{K}\mathcal{L}(\boldsymbol{\alpha}_2, 1) + \frac{1}{K}\mathcal{L}(\boldsymbol{\alpha}_3, 1).$$

By analogy, we can obtain

$$\mathcal{L}_2 = b_1\mathcal{L}(\boldsymbol{\alpha}_1, 1) + b_2\mathcal{L}(\boldsymbol{\alpha}_1, 1) + \cdots + b_K\mathcal{L}(\boldsymbol{\alpha}_K, 1) \leq \frac{1}{K}\mathcal{L}(\boldsymbol{\alpha}_1, 1) + \frac{1}{K}\mathcal{L}(\boldsymbol{\alpha}_2, 1) + \cdots + \frac{1}{K}\mathcal{L}(\boldsymbol{\alpha}_K, 1) = \mathcal{L}_1$$

□

✍ **Additional explanation:** The condition given in Proposition A.2, $w_1 \geq w_2 \geq \cdots \geq w_K$ for $\mathcal{L}(\boldsymbol{\alpha}_1, 1) \leq \mathcal{L}(\boldsymbol{\alpha}_2, 1) \leq \cdots \leq \mathcal{L}(\boldsymbol{\alpha}_K, 1)$, states that there is a higher weight for the instance whose prediction is closer to the expected evidence derived from positive instances. Here, we assume that the instance-level estimator $T = g \circ f$ could predict a higher w_k for positive instances and a lower one for negative instances, in order to satisfy that condition approximately.

With Proposition A.1 and Proposition A.2, we give the proof of Proposition 4.3 as follows:

***Proposition** 4.3*. Let $\mathcal{L}(\boldsymbol{\alpha}, y)$ be a loss function *w.r.t* $\boldsymbol{\alpha}$ and y. For any positive bag $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_K\}$, assume $\bar{w}_k \geq 0\ \forall k \in [1, K]$, $\sum_k \bar{w}_k = 1$, and $\widetilde{\boldsymbol{\alpha}} = \sum_k \bar{w}_k \boldsymbol{\alpha}_k$. $\mathcal{L}_{\text{ins}}^{+} = \mathcal{L}(\widetilde{\boldsymbol{\alpha}}, 1) \leq \sum_k \bar{w}_k \mathcal{L}(\boldsymbol{\alpha}_k, 1) \leq \sum_k \frac{1}{K}\mathcal{L}(\boldsymbol{\alpha}_k, 1)$ holds in instance evidential learning, when $\mathcal{L}$ is a convex function *w.r.t* $\boldsymbol{\alpha}$ and there is $\bar{w}_1 \geq \bar{w}_2 \geq \cdots \geq \bar{w}_K$ for $\mathcal{L}(\boldsymbol{\alpha}_1, 1) \leq \mathcal{L}(\boldsymbol{\alpha}_2, 1) \leq \cdots \leq \mathcal{L}(\boldsymbol{\alpha}_K, 1)$.

Proof. Since $\sum_k \bar{w}_k \mathcal{L}(\boldsymbol{\alpha}_k, 1) = \mathcal{L}_2$ and $\mathcal{L}$ is a convex function *w.r.t* $\boldsymbol{\alpha}$, we have $\mathcal{L}_{\text{ins}}^{+} \leq \mathcal{L}_2$ according to Proposition A.1. Further, $\sum_k \frac{1}{K}\mathcal{L}(\boldsymbol{\alpha}_k, 1) = \mathcal{L}_1$ and those given conditions exactly satisfy the condition of Proposition A.2, so $\mathcal{L}_2 \leq \mathcal{L}_1$. Hence, there is $\mathcal{L}_{\text{ins}}^{+} \leq \mathcal{L}_2 \leq \mathcal{L}_1$. Namely, $\mathcal{L}_{\text{ins}}^{+} = \mathcal{L}(\widetilde{\boldsymbol{\alpha}}, 1) \leq \sum_k \bar{w}_k \mathcal{L}(\boldsymbol{\alpha}_k, 1) \leq \sum_k \frac{1}{K}\mathcal{L}(\boldsymbol{\alpha}_k, 1)$ holds. □

Proposition 4.3 ensure that our objective function $\mathcal{L}_{\text{ins}}^{+}$ can provide a tighter upper bound than $\mathcal{L}_1$ and $\mathcal{L}_2$ for the ideal objective function Eq.(14) under given conditions. This implies that our optimization strategy (3) could yield a more suitable weakly-supervised posterior $\hat{\boldsymbol{\theta}}_{\mathbf{w}}$ in instance evidential learning, such that $p(\boldsymbol{\theta}_{\mathbf{w}}|\mathcal{D}) \approx \delta(\boldsymbol{\theta}_{\mathbf{w}} - \hat{\boldsymbol{\theta}}_{\mathbf{w}})$. Accordingly, we could approximate the intractable posterior in Eq.(4) with $p(\boldsymbol{\nu}|\mathbf{x}^*, \mathcal{D}) \approx p(\boldsymbol{\nu}|\mathbf{x}^*, \hat{\boldsymbol{\theta}}_{\mathbf{w}})$ for a new instance input $\mathbf{x}^*$, as follows:

$$P(y^*|\mathbf{x}^*, \mathcal{D}) = \int P(y^*|\mathbf{x}^*, \boldsymbol{\theta}_{\mathbf{w}})p(\boldsymbol{\theta}_{\mathbf{w}}|\mathcal{D})d\boldsymbol{\theta}_{\mathbf{w}} = \int P(y^*|\boldsymbol{\nu})p(\boldsymbol{\nu}|\mathbf{x}^*, \mathcal{D})d\boldsymbol{\nu} \approx \int P(y^*|\boldsymbol{\nu})p(\boldsymbol{\nu}|\mathbf{x}^*, \hat{\boldsymbol{\theta}}_{\mathbf{w}})d\boldsymbol{\nu},$$

where $\boldsymbol{\nu}$ is instance-level predictive probability. The above equation with Dirichlet distribution has a closed-form solution for instance-level uncertainty quantification, as elaborated in Appendix B.2. The ablation studies on instance loss functions (refer to Appendix E.2 and Appendix F.3) empirically demonstrate Proposition 4.3.

B. Uncertainty Measures

This section mainly shows common measures for uncertainty quantification. Two predictive distributions, general Categorical distribution and related Dirichlet distribution, are considered here. This section is adapted from Malinin & Gales (2018), in order to provide readers with additional reference.

B.1. Measures for Predictive Categorical Distribution

Given a predictive Categorical distribution for input X, $P(Y|X)$, its *total uncertainty* could be quantified through two common measures as follows:

(1) Max probability It is the probability of the predicted category, as a measure of confidence in prediction. Intuitively, a larger max probability means that model is more confident of its prediction. Max probability can be written as follows:

$$\max_i P(Y=i|X),$$

where $i \in [1, C]$ and C is the total number of categories.

(2) Entropy By definition, it is calculated by

$$\mathcal{H}[P(Y|X)] = -\sum_{i=1}^{C} \big[P(Y=i|X) \ln P(Y=i|X)\big]$$

A flat predictive distribution would yield a maximum $\mathcal{H}[P(Y|X)]$, implying high predictive uncertainty.

Moreover, from the perspective of Bayesian (Neal, 2012), consider the posterior distribution of model parameter given dataset $\mathcal{D}$, *i.e.*, $p(\boldsymbol{\omega}|\mathcal{D})$. There is another measure commonly used for quantifying the *model uncertainty* in prediction:

$$\underbrace{I[Y,\boldsymbol{\omega}|X,\mathcal{D}]}_{\text{Model Uncertainty}} = \underbrace{\mathcal{H}\big[\mathbb{E}_{p(\boldsymbol{\omega}|\mathcal{D})}P(Y|X,\boldsymbol{\omega})\big]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{p(\boldsymbol{\omega}|\mathcal{D})}\big[\mathcal{H}[P(Y|X,\boldsymbol{\omega})]\big]}_{\text{Expected Data Uncertainty}}. \tag{16}$$

It is referred to as **Mutual information** (MI). As shown in the equation above, MI can be cast as the difference of *total uncertainty* and *expected data uncertainty*. The former is captured by $\mathcal{H}\big[\mathbb{E}_{p(\boldsymbol{\omega}|\mathcal{D})}P(Y|X,\boldsymbol{\omega})\big]$, *i.e.*, the entropy of expected predictive distribution. The latter is captured by $\mathbb{E}_{p(\boldsymbol{\omega}|\mathcal{D})}\big[\mathcal{H}[P(Y|X,\boldsymbol{\omega})]\big]$, *i.e.*, the expected entropy of predictive distribution. For traditional non-Bayesian NN models, MI is zero because their parameter is usually a point estimation.

B.2. Measures for Predictive Dirichlet Distribution

Consider a prediction with Dirichlet distribution (known as the conjugate prior of Categorical distribution):

$$Dir(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{C}\Gamma(\alpha_i)} \prod_{i=1}^{C} p_i^{\alpha_i - 1},$$

where $\boldsymbol{p} \in \mathcal{S}^{C-1}$ (a probability simplex with $C-1$ dimensions), $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_C]$, $\alpha_i \geq 0\ \forall i \in [1, C]$, $\Gamma(\cdot)$ is a *gamma* function, and $\alpha_0 = \sum_{i=1}^{C} \alpha_i$ often called the precision or Dirichlet strength. Its expected probability is as follows:

$$\mathbb{E}[\boldsymbol{p}] = \Big[\frac{\alpha_1}{\alpha_0}, \frac{\alpha_2}{\alpha_0}, \cdots, \frac{\alpha_C}{\alpha_0}\Big].$$

Next, we give common uncertainty measures for $Dir(\boldsymbol{p}|\boldsymbol{\alpha})$. All of them have a closed-form analytical solution.

Max probability and Entropy By simply taking $\mathbb{E}[\boldsymbol{p}]$ as prediction, they can be written as

$$\max\Big\{\frac{\alpha_i}{\alpha_0}\Big\}_{i=1}^{C} \text{ and } \mathcal{H}\big[\mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha})}P(Y|\boldsymbol{p})\big] = -\sum_{i=1}^{C}\frac{\alpha_i}{\alpha_0}\ln\frac{\alpha_i}{\alpha_0},$$

respectively, to capture the *total uncertainty* in prediction.

Expected Entropy Different from the calculation of predictive entropy above, expected entropy is expressed as

$$\mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha})}\big[\mathcal{H}[P(Y|\boldsymbol{p})]\big] = \int_{\mathcal{S}^{C-1}} Dir(\boldsymbol{p}|\boldsymbol{\alpha})\Big(-\sum_{i=1}^{C} p_i \ln p_i\Big)d\boldsymbol{p} = -\sum_{i=1}^{C}\frac{\alpha_i}{\alpha_0}\big(\psi(\alpha_i+1) - \psi(\alpha_0+1)\big),$$

where $\psi(\cdot)$ is a *digamma* function defined as $\psi(x) = \frac{d}{dx}\log\Gamma(x)$. As mentioned in Eq.(16), expected entropy is usually utilized to measure *data uncertainty*. Intuitively, it could capture the 'peak' probabilities in $\mathbb{E}[\boldsymbol{p}]$.

Mutual Information By definition, mutual information (MI) can be written as follows:

$$\underbrace{I[Y,\boldsymbol{p}|X,\mathcal{D}]}_{\text{Distributional Uncertainty}} = \underbrace{\mathcal{H}\big[\mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{p}|X,\mathcal{D})}P(Y|\boldsymbol{p})\big]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{p}|X,\mathcal{D})}\big[\mathcal{H}[P(Y|\boldsymbol{p})]\big]}_{\text{Expected Data Uncertainty}}.$$

This equation calculates the MI between Y and the categorical $\boldsymbol{p}$, rather than the $\boldsymbol{\omega}$ written in Eq.(16). Thereby, $I[Y, \boldsymbol{p}|X, \mathcal{D}]$ is generally used to measure *distributional uncertainty*. Assuming there is a sufficient point estimate $\hat{\boldsymbol{\omega}}$ satisfying $Dir(\boldsymbol{p}|X, \mathcal{D}) \approx Dir(\boldsymbol{p}|X, \hat{\boldsymbol{\omega}}) = Dir(\boldsymbol{p}|\boldsymbol{\alpha})$ where $\boldsymbol{\alpha}$ is the prediction of X given the model parameter $\hat{\boldsymbol{\omega}}$, MI (Malinin & Gales, 2018) could be approximated and calculated as follows:

$$\begin{aligned}
\underbrace{I[Y, \boldsymbol{p}|X, \mathcal{D}]}_{\text{Distributional Uncertainty}} &\approx \underbrace{\mathcal{H}\big[\mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha})} P(Y|\boldsymbol{p})\big]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{\boldsymbol{p}\sim Dir(\boldsymbol{\alpha})}\big[\mathcal{H}[P(Y|\boldsymbol{p})]\big]}_{\text{Expected Data Uncertainty}} \\
&= -\sum_{i=1}^{C} \frac{\alpha_i}{\alpha_0} \ln \frac{\alpha_i}{\alpha_0} - \left(-\sum_{i=1}^{C} \frac{\alpha_i}{\alpha_0} \big(\psi(\alpha_i + 1) - \psi(\alpha_0 + 1)\big) \right) \\
&= -\sum_{i=1}^{C} \frac{\alpha_i}{\alpha_0} \Big(\ln \frac{\alpha_i}{\alpha_0} - \psi(\alpha_i + 1) + \psi(\alpha_0 + 1) \Big).
\end{aligned} \tag{17}$$

More measures involving concentration parameters $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_C]$, *e.g.*, $\max_i\{\alpha_i\}$, $\alpha_0 = \sum_{i=1}^{C} \alpha_i$, and $\frac{C}{\sum_{i=1}^{C} \alpha_i}$, are often utilized in Dirichlet-based uncertainty (DBU) models, to capture model uncertainty, as $\boldsymbol{\alpha}$ shapes the distribution of predictive probabilities. For more discussions about this, readers could refer to Ulmer et al. (2023).

C. Objective Function for Evidential Deep Learning

For completeness, here we provide the details of Fisher Information-based objective function (Deng et al., 2023). This section is adapted from Deng et al. (2023).

The Fisher Information-based objective function employed by us, $\mathcal{L}_{\mathcal{I}\text{-EDL}}$, can be written as follows:

$$\begin{aligned}
\min \ & \mathbb{E}_{(X,Y)\sim\mathcal{P}} \mathbb{E}_{\boldsymbol{\mu}\sim Dir(\boldsymbol{\alpha})} \big[-\log p(Y|\boldsymbol{\mu}, \boldsymbol{\alpha}, \sigma^2) \big]. \\
s.t. \ & \boldsymbol{\alpha} = \mathcal{F}(X) + 1 \\
& p(Y|\boldsymbol{\mu}, \boldsymbol{\alpha}, \sigma^2) = \mathcal{N}\big(Y|\boldsymbol{\mu}, \sigma^2 \mathcal{I}(\boldsymbol{\alpha})^{-1}\big) \\
& \mathcal{I}(\boldsymbol{\alpha}) = \mathbb{E}_{Dir(\boldsymbol{\mu}|\boldsymbol{\alpha})} \Big[-\frac{\partial^2 \log Dir(\boldsymbol{\mu}|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \boldsymbol{\alpha}^T} \Big]
\end{aligned} \tag{18}$$

In this function, $\boldsymbol{\alpha}$ is the concentration parameter of Dirichlet distribution predicted from X by $\mathcal{F}(\cdot)$. Moreover, $p(Y|\boldsymbol{\mu}, \boldsymbol{\alpha}, \sigma^2)$ is assumed to be a multivariate Gaussian distribution $\mathcal{N}\big(Y|\boldsymbol{\mu}, \sigma^2 \mathcal{I}(\boldsymbol{\alpha})^{-1}\big)$, and $\mathcal{I}(\boldsymbol{\alpha})$ is referred to as Fisher Information Matrix (**FIM**) for $Dir(\boldsymbol{\alpha})$.

Intuitively, FIM is introduced into the variance of predictive distribution to obtain a non-isotropic Normal distribution for the label generation of a specific sample. As a result, there is an adaptive weight provided by FIM. This weight is assigned to each class in eventual loss function, to adjust the information of each class contained in the sample. This could avoid the potential over-penalty of some classes in the supervision based on one-hot labels.

Given a dataset $\{(X_j, Y_j)\}_{j=1}^{N}$, the eventual form of $\mathcal{L}_{\mathcal{I}\text{-EDL}}$ is as follows:

$$\min \frac{1}{N} \sum_{j=1}^{N} \big(\mathcal{L}_j^{\mathcal{I}\text{-MSE}} - \lambda_1 \mathcal{L}_j^{|\mathcal{I}|} + \lambda_2 \mathcal{L}_j^{\text{KL}} \big),$$

where

$$\mathcal{L}_j^{\mathcal{I}\text{-MSE}} = \sum_{i=1}^{C} \left((y_{ji} - \frac{\alpha_{ji}}{\alpha_{j0}})^2 + \frac{\alpha_{ji}(\alpha_{j0} - \alpha_{ji})}{\alpha_{j0}^2(\alpha_{j0} + 1)} \right) \psi^{(1)}(\alpha_{ji}),$$

$$\mathcal{L}_j^{|\mathcal{I}|} = \sum_{i=1}^{C} \log \psi^{(1)}(\alpha_{ji}) + \log \left(1 - \sum_{i=1}^{C} \frac{\psi^{(1)}(\alpha_{j0})}{\psi^{(1)}(\alpha_{ji})} \right),$$

$$\mathcal{L}_j^{\text{KL}} = \log \Gamma\Big(\sum_{i=1}^{C} \hat{\alpha}_{ji} \Big) - \log \Gamma(C) - \sum_{i=1}^{C} \log \Gamma(\hat{\alpha}_{ji}) + \sum_{i=1}^{C} (\hat{\alpha}_{ji} - 1) \left[\psi(\hat{\alpha}_{ji}) - \psi\Big(\sum_{c=1}^{C} \hat{\alpha}_{jc} \Big) \right],$$

and $\lambda_1 \geq 0, \lambda_2 \geq 0$. Moreover, $\hat{\alpha}_{ji} = \alpha_{ji}(1 - Y_{ji}) + Y_{ji}\ \forall j \in [1, N], i \in [1, C]$. $\psi(\cdot)$ is a *digamma* function defined as $\psi(x) = \frac{d}{dx}\log\Gamma(x)$, $\psi^{(1)}(\cdot)$ is a *trigamma* function with $\psi^{(1)}(x) = \frac{d}{dx}\psi(x)$, and $\Gamma(\cdot)$ stands for a *gamma* function. Readers could refer to Deng et al. (2023) for derivation details.

For better understanding, we briefly explain these terms as follows:

(1) For the first term $\mathcal{L}_j^{\mathcal{I}\text{-MSE}}$, it introduces a new $\psi^{(1)}(\alpha_{ji})$ to the MSE (mean square error) loss $\mathcal{L}_j^{\text{MSE}}$ frequently-used in EDL (Sensoy et al., 2018). Concretely,

$$\mathcal{L}_j^{\text{MSE}} = \sum_{i=1}^{C}\left((y_{ji} - \frac{\alpha_{ji}}{\alpha_{j0}})^2 + \frac{\alpha_{ji}(\alpha_{j0} - \alpha_{ji})}{\alpha_{j0}^2(\alpha_{j0} + 1)}\right).$$

It is derived from a simple MSE-based Bayes risk function:

$$\mathcal{L}_{\text{MSE}} = \int ||Y - \boldsymbol{\mu}||_2^2 Dir(\boldsymbol{\mu}|\boldsymbol{\alpha})d\boldsymbol{\mu}.$$

$\psi^{(1)}(\alpha_{ji})$ is specially added into $\mathcal{L}_j^{\text{MSE}}$, in order to encourage the model to focus more on the class with low evidence.

(2) For the second term $-\mathcal{L}_j^{|\mathcal{I}|}$, it is equal to $-\log|\mathcal{I}(\boldsymbol{\alpha}_j)|$, *i.e.*,

$$-\mathcal{L}_j^{|\mathcal{I}|} = -\log|\mathcal{I}(\boldsymbol{\alpha}_j)|.$$

It is taken to avoid the overconfidence caused by excessive evidence.

(3) For the final term $\mathcal{L}_j^{\text{KL}}$, its original form (Sensoy et al., 2018) is as follows:

$$\mathcal{L}_j^{\text{KL}} = \text{KL}(Dir(\boldsymbol{\mu}|\hat{\boldsymbol{\alpha}}_j)||Dir(\boldsymbol{\mu}|\mathbf{1})),$$

where $\text{KL}(\cdot)$ is a function measuring Kullback-Leibler (KL) divergence. Moreover, $\hat{\boldsymbol{\alpha}}_j = \boldsymbol{\alpha}_j \odot (1 - \boldsymbol{Y}_j) + \boldsymbol{Y}_j$ where $\boldsymbol{Y}_j$ stands for the one-hot label of Y_j. It indicates manually masking the predicted parameter corresponding to the ground-truth class. Therefore, $\mathcal{L}_j^{\text{KL}}$ can be view as a loss term aiming to suppress the evidence of irrelevant classes.

D. Experimental Details

This section provides the additional details of experimental setup (Section 6.1), including bag generation (D.1), datasets (D.2), and implementation and network training (D.3). Our source code has been submitted as Supplementary Material.

D.1. Bag Generation

Following Ilse et al. (2018), we generate a bag dataset for MIL from a given single-instance dataset. **(1) Steps**: At first, we set **a class of interest** (as positive class) and this class is from the given dataset. Then, we randomly select a certain number of samples from the given dataset to form a multi-instance bag. This bag is positive if it contains at least one sample from the class of interest; otherwise, it is negative. Bag length follows a Normal distribution $\mathcal{N}(10, 2)$. **(2) More settings**: positive and negative bags are generated sequentially in a loop to obtain a balanced bag dataset. The ratio of positive instances roughly follows an Uniform distribution $\text{U}(0, 1)$ for positive bags. No that, we ensure that all the instances of training bags are only sampled from the training set of the given dataset, and do so for validation and test bags.

D.2. Datasets

MNIST-bags (LeCun, 1998) Following the dataset setting in Ilse et al. (2018), there are 500, 100, and 1000 generated MNIST bags in training, validation, and test set, respectively. Each bag contains multiple MNIST images (each image with the size of $1 \times 28 \times 28$). The number '9' is set as the class of interest, as it is easily mistaken with '7' and '4' in hand-written numbers. In OOD detection tasks, **FMNIST-bags** (Xiao et al., 2017) and **KMNIST-bags** (Clanuwat et al., 2018) are taken as OOD MIL datasets. Both of two contain 1000 OOD bags. The length of these OOD bags also follows $\mathcal{N}(10, 2)$. We show bag examples in Fig. 4.

CIFAR10-bags (Krizhevsky et al., 2009) We generate 6000, 1000 and 2000 CIFAR bags for training, validation, and test, respectively. This dataset is more complex than MNIST-bags since its instances, CIFAR10 images, are more diverse than

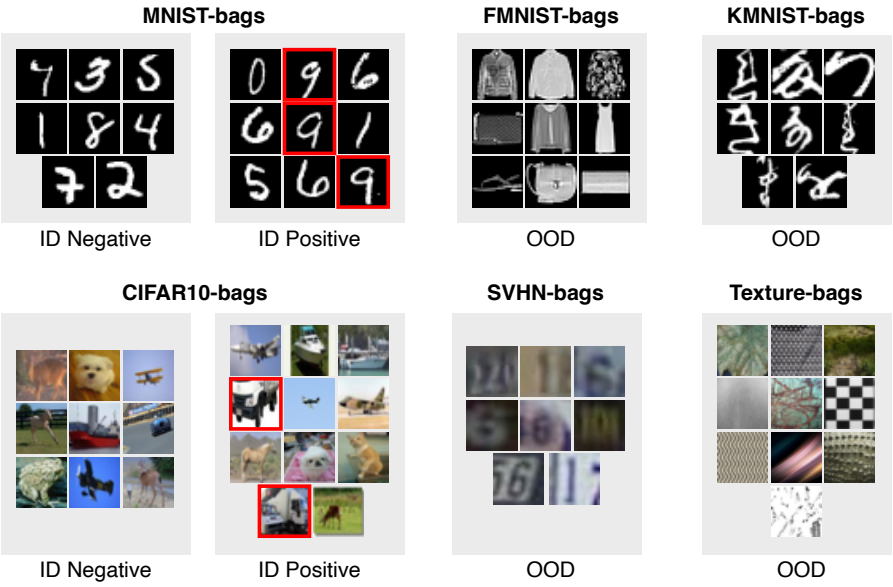

Figure 4. Bag examples of each bag dataset. Red box indicates the class of interest, '9' and 'truck' for MNIST and CIFAR10, respectively.

MNIST ones. A single instance (image) in each bag is with the size of $3 \times 32 \times 32$. We randomly select **'truck'** as the class of interest. In OOD detection tasks, **SVHN-bags** (Yuval, 2011) and **Texture-bags** (Cimpoi et al., 2014) are two OOD MIL datasets; and each contains 2,000 OOD bags. OOD instances are resized to $3 \times 32 \times 32$, in order to keep the same size as CIFAR10 instances. Similarly, the length of OOD bags follows $\mathcal{N}(10, 2)$. Bag examples are exhibited in Fig. 4.

CAMELYON16 (Bejnordi et al., 2017) It is a real-world pathology dataset, originally proposed for breast cancer lymph node metastasis detection and frequently used for evaluating MIL algorithms. We obtain 270 and 129 histopathology WSIs (Whole-Slide Images) for training and test, respectively, provided by official organizers. There are 111 tumor slides and 159 normal slides in the training set, and 49 tumor slides and 80 normal slides in the test set. We leave 15% training samples as a validation set. Please refer to Fig. 5(a) for WSI examples. More details of CAMELYON16 are as follows:

- **Preprocessing**: Since a single WSI has extremely-high resolution (*e.g.*, $40,000 \times 40,000$ pixels), we process each image into a bag of feature vectors with CLAM (Lu et al., 2021) by three steps: i) tissue region selection, ii) image patching, and iii) patch feature extraction. Each patch is an image with 256×256 pixels from the WSI at $20\times$ magnification. Feature vector is extracted from patch image by a fixed (frozen) deep network. This fixed network is pre-trained on the patches of *training samples* by self-supervised learning, provided by Li et al. (2021). As a result, there are 11,753 instances in each WSI bag on average, and each instance is a feature vector with the length of 512.
- **OOD dataset**: The histopathology WSIs of *prostate cancer* are used as the OOD samples of CAMELYON16 (breast cancer), following Linmans et al. (2023). These WSIs are from **TCGA-PRAD** (The Cancer Genome Atlas Prostate Adenocarcinoma[1]) (Kandoth et al., 2013), containing 449 diagnostic images. Their preprocessing is the same as that of CAMELYON16. Finally, there are 3,484 instances in each bag on average. TCGA-PRAD samples are shown in Fig. 5(a). They often present differences with CAMELYON16 in cell distribution and tissue morphology.
- **Distribution shift dataset**: Given the test WSIs of CAMELYON16, we synthesize its three shifted versions using the image noises with different strengths, called *lighter*, *light*, and *strong*. Specifically, Gaussian Blurring or HED (Hematoxylin-Eosin-DAB) color variation is applied to the patch images of each test WSI, to simulate the possible noises in digital pathology, following Tellez et al. (2019) and Liu et al. (2024a). The patch image samples with different noises are shown in Fig. 5(b). Eventually, all the patch images with noises are transformed into instances (feature vectors) using the same feature extractor mentioned in WSI preprocessing.

[1]Available at https://portal.gdc.cancer.gov/projects/TCGA-PRAD

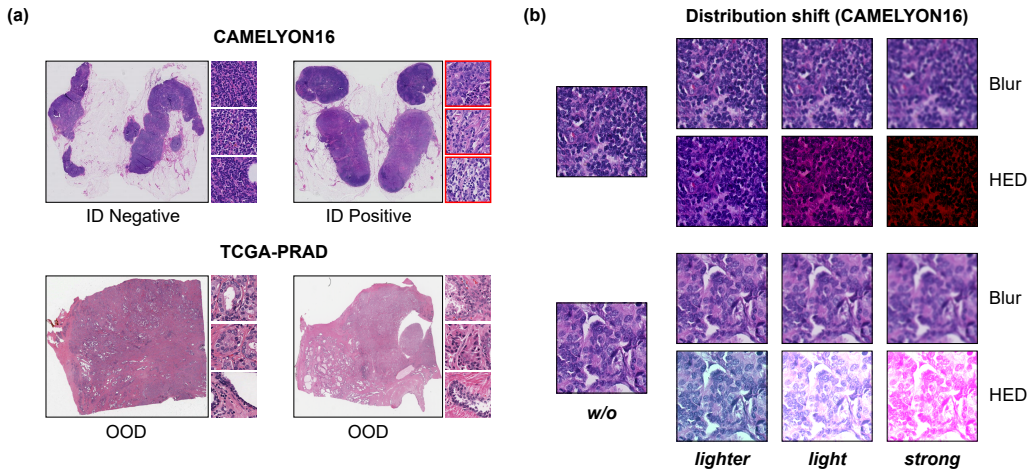

Figure 5. (a) Samples of CAMELYON16 and TCGA-PRAD. Red box indicates the patch with tumorous cells. (b) Patch samples of distribution shift CAMELYON16. Blur and HED mean the image noise of Gaussian Blurring and HED color variation, respectively.

D.3. Implementation and Network Training

Deep MIL networks The most representative deep MIL networks, *e.g.*, Mean, Max, ABMIL (Ilse et al., 2018), and DSMIL (Li et al., 2021), are adopted in our experiments. From a unified perspective, these networks comprise three key parts as follows. (1) **Instance encoder**. We employ LeNet (LeCun, 1998) as the encoder to transform MNIST images into instance embeddings, following Ilse et al. (2018). For CIFAR10-bags, a modified AlexNet (Krizhevsky et al., 2012) is adopted. For CAMELYON16, we directly use an MLP (Multi-Layer Perceptron) layer, since image patches have been transformed into feature vectors. (2) **MIL pooling operator**. Mean and max-based pooling are used for Mean and Max MIL networks, respectively. For ABMIL and DSMIL, we follow their respective implementation in MIL pooling. Specifically, for ABMIL, a standard attention mechanism, rather than its gated variant, is adopted in MIL pooling, because it is more efficient in computation and is often competitive in performance, compared to its gated variant (Ilse et al., 2018; Shi et al., 2020). (3) **Classification head**. It is a fully-connected layer with negative and positive output nodes.

Related UE methods For the classical UE method, **Deep Ensemble** (Lakshminarayanan et al., 2017), by default we train 10 ABMIL networks with different random seeds. For **MC Dropout** (Gal & Ghahramani, 2016), Dropout layers are used in the instance encoder of ABMIL, with a drop rate of 0.25; 10 estimates are sampled from the network and their mean is taken as prediction. For BNN-based **Bayes-MIL** (Cui et al., 2023), we follow its implementation to sample 16 estimates. Lastly, for **$\mathcal{I}$-EDL** (Deng et al., 2023), we modify the classification head of ABMIL into an evidential output layer (Sensoy et al., 2018), and adopt a $\mathcal{I}$-EDL loss function for evidential learning.

MIREL Its implementation details are as follows. **(1) Bag-level network:** Our method could be combined with existing deep MIL networks for MIUE. Thus, we directly follow their implementations and employ them to implement our bag-level networks. In particular, we replace their conventional classification head with an evidential output layer (Sensoy et al., 2018), and utilize $\exp(\cdot)$ for the implementation of $\mathcal{A}(\cdot)$. **(2) instance-level network:** For our proposed residual instance estimator, $R(\mathbf{x}) = T(\mathbf{x}) + r_\pi(f_\psi(\mathbf{x}))$, $T(\cdot)$ is exactly the bag-level network, $f_\psi(\cdot)$ is the instance encoder of bag-level network, and $r_\pi(\cdot)$ is simply implemented by an MLP layer. To make instance-level evidential learning more stable, we adopt $\tanh(\cdot)$ to let r_π output a scale value in $[-1, 1]$. This scale value expresses a residual estimate proportional to $T(\mathbf{x})$. **(3) Loss function:** Apart from Fisher Information-based objective function (Appendix C), a RED loss (Pandey & Yu, 2023) is also adopted to avoid zero-evidence regions in EDL, as stated in the last paragraph of Section 4.3. **(4) Optimization strategy for $\mathcal{L}_{\text{MIREL}}$:** In the experiments on MNIST-bags and CAMELYON16, the bag-level parameter ψ and the instance-level parameter π are optimized by $\mathcal{L}_{\text{MIREL}}$ in weakly-supervised instance residual evidential learning. While on CIFAR10-bags, only the instance-level parameter π is involved in the optimization of $\mathcal{L}_{\text{MIREL}}$. Namely, we specially freeze the bag-level parameter ψ in optimizing $\mathcal{L}_{\text{MIREL}}$. We will elaborate on this setting in Appendix F.3. **(5) Hyper-parameters:** Following Deng et al. (2023), the coefficient λ_1 of $-\mathcal{L}_j^{|\mathcal{I}|}$ is set by a grid-search over (0.1, 0.05, 0.01, 0.005, 0.001), and the coefficient λ_2 of $\mathcal{L}_j^{\text{KL}}$

is set to $\min(1, \frac{t}{10}) \in [0, 1]$, where t is the index of current training epoch.

Network training Learning rate, by default, is set to 0.0001 and it decays by a factor of 0.5 when the criterion on validation set does not decrease within 10 epochs. The other default settings are as follows: an epoch number of 200, a batch size of 1 (bag), a gradient accumulation step of 8, and an optimizer of Adam with a weight decay rate of 1×10^{-5}. Early stopping is applied when the criterion on validation set does not decrease within 20 epochs by default. The sum of loss and error is adopted as the criterion. Moreover, EDL-based models, *e.g.*, $\mathcal{I}$-EDL and our MIREL, are trained using the same $\mathcal{L}_{\mathcal{I}\text{-EDL}}$; while the other standard classification models use $\mathcal{L}_{\text{BCE}}$, *i.e.*, a BCE (binary cross-entropy) loss. In ablation study, three types of models are trained. Their details are shown in Table 4.

Table 4. Details of the models used in ablation study.

Network	Loss	Instance prediction
Deep MIL	$\mathcal{L}_{\text{BCE}}$	a_k (attention score) $T(\mathbf{x})$
Deep MIL + EDL	$\mathcal{L}_{\mathcal{I}\text{-EDL}} + \mathcal{L}_{\text{RED}}$	a_k (attention score) $T(\mathbf{x})$
Deep MIL + MIREL	$\mathcal{L}_{\mathcal{I}\text{-EDL}} + \mathcal{L}_{\text{MIREL}} + \mathcal{L}_{\text{RED}}$	$R(\mathbf{x})$

E. Additional Results on MNIST-bags

E.1. Uncertainty Analysis

ABMIL (1) The result of uncertainty analysis on KMNIST-bags is shown in Fig. 6. From this result, we could see that i) the ABMIL w/ MIREL can provide more discriminative predictive uncertainty for the bags with different OOD ratios, compared to original ABMIL; ii) our method can assist ABMIL in distinguishing ID (MNIST) instances and OOD (KMNIST) ones more accurately. (2) α_0 is another commonly-used uncertainty measure for EDL models. Its distribution at instance and bag levels are shown in Fig. 7. These results suggest that our MIREL could also detect OOD samples through the concentration parameter α_0.

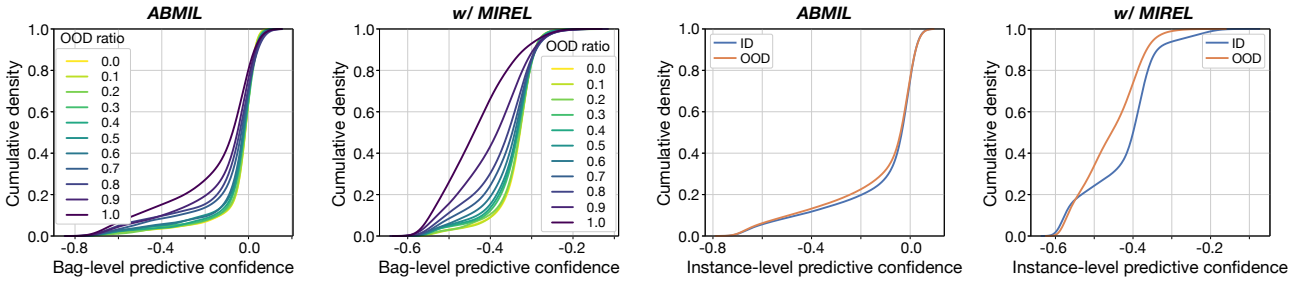

Figure 6. Distribution of bag-level and instance-level predictive confidence (negative expected entropy). **MNIST-bags** is ID dataset. The OOD instances used in this experiment are from KMNIST.

DSMIL We show more results of uncertainty analysis, in which DSMIL is taken as the base MIL network. These results contain bag-level UE (Fig. 8), instance-level UE (Fig. 9), and α_0 estimate (Fig. 10). We summarize our observations from these as follows. (1) When OOD dataset is FMNIST-bags, our MIREL helps DSMIL to provide more accurate uncertainty for the bags with different OOD ratios, while vanilla DSMIL often shows overconfident prediction and cannot response to abnormal bags. When using the bags with different OOD instance ratios, there is no obvious change in bag-level predictive confidence for both DSMIL and its MIREL counterpart. (2) At instance level, DSMIL often mistakenly predicts more confidence for OOD instances than ID ones. After combining with our MIREL, DSMIL tends to assign ID instances with more confidence. (3) There are similar findings from the results of α_0 for the DSMIL models with our MIREL.

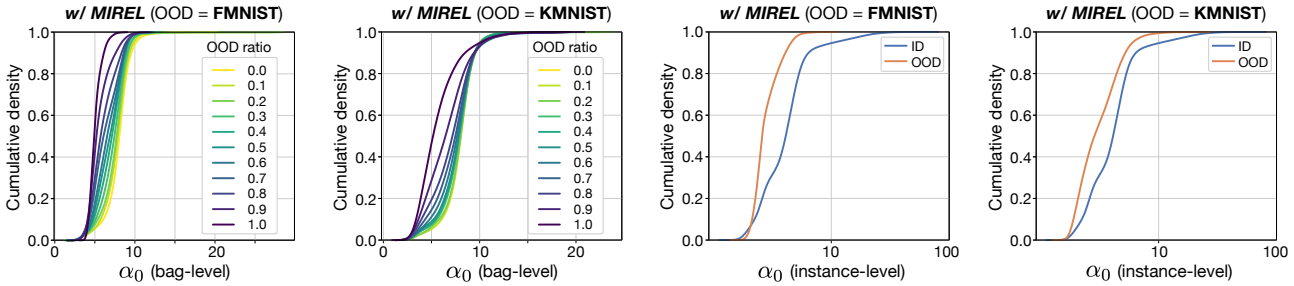

Figure 7. Distribution of bag-level and instance-level α_0 output by the ABMIL models with our MIREL. **MNIST-bags** is ID dataset.

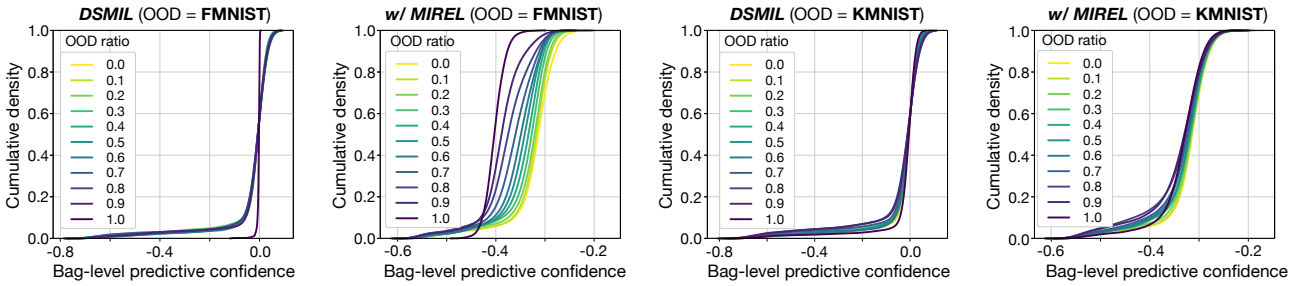

Figure 8. Distribution of bag-level predictive confidence (negative expected entropy). DSMIL is the base MIL network in this experiment. ID dataset is **MNIST-bags**.

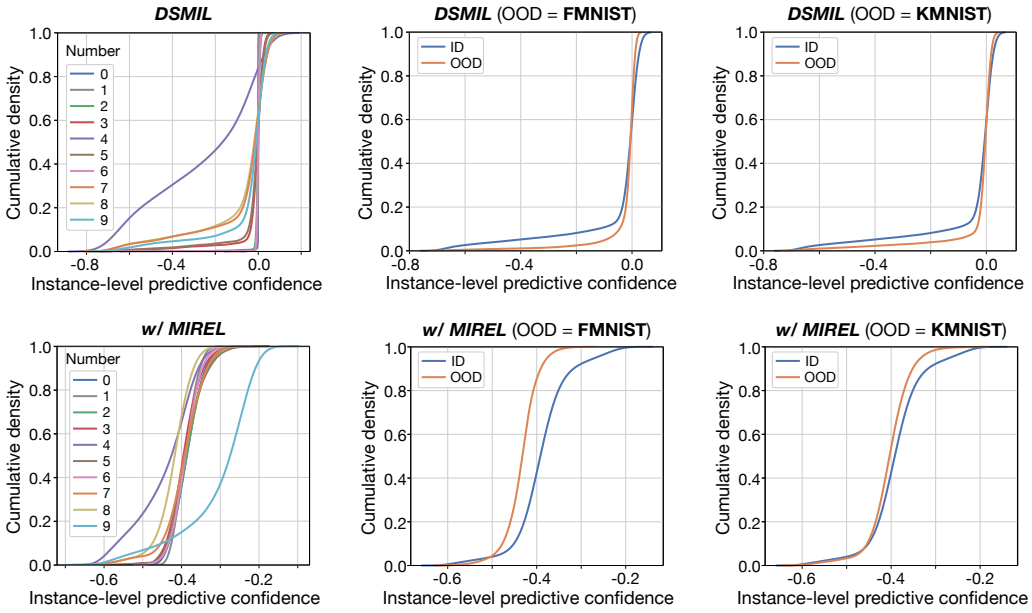

Figure 9. Distribution of instance-level predictive confidence (negative expected entropy). DSMIL is the base MIL network in this experiment. ID dataset is **MNIST-bags**.

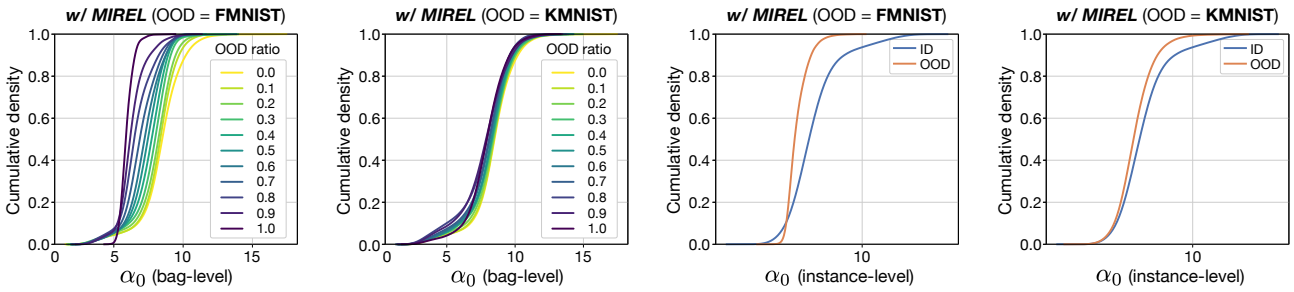

Figure 10. Distribution of bag-level and instance-level α_0 output by the DSMIL models with our MIREL. ID dataset is **MNIST-bags**.

E.2. Ablation Study

Optimization strategy for $R(\mathbf{x})$ We adopt three different optimization strategies for *the instances from positive bags*. They can be represented by three loss functions, $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_{\text{ins}}^{+}$, as described in Appendix A.3. Test results are shown in Table 5. Our main findings are as follows. (1) Compared to $\mathcal{L}_{\text{ins}}^{+}$, the UE performance obtained by $\mathcal{L}_1$ often lags far behind. (2) Compared to $\mathcal{L}_{\text{ins}}^{+}$, $\mathcal{L}_2$ leads to the worse overall UE performance at bag level, with a drop of 2.84%, although it performs slightly better at instance level, with a narrow increase of 0.77%. These findings empirically demonstrate the effectiveness of our weakly-supervised evidential learning strategy.

Table 5. Ablation study on the loss function used for training $R(\mathbf{x})$. The base MIL network is ABMIL and it is trained on **MNIST-bags**.

Loss	Bag-level					Instance-level				
	Acc.	Conf.	OOD-F	OOD-K	$\overline{UE}$	Acc.	Conf.	OOD-F	OOD-K	$\overline{UE}$
$\mathcal{L}_1$	96.46 $\pm$ 0.59	83.30 $\pm$ 4.43	90.91 $\pm$ 2.46	76.89 $\pm$ 2.76	83.70	85.70 $\pm$ 0.50	85.42 $\pm$ 5.48	66.63 $\pm$ 4.77	63.00 $\pm$ 4.82	71.68
$\mathcal{L}_2$	96.46 $\pm$ 0.30	84.40 $\pm$ 2.41	91.15 $\pm$ 2.91	75.33 $\pm$ 3.14	83.63	86.29 $\pm$ 0.59	90.67 $\pm$ 1.71	**81.27** $\pm$ 3.63	66.26 $\pm$ 3.53	**79.40**
$\mathcal{L}_{\text{ins}}^{+}$	96.48 $\pm$ 0.22	**86.63** $\pm$ 1.32	**92.84** $\pm$ 0.60	**79.95** $\pm$ 4.12	**86.47**	87.71 $\pm$ 0.67	**90.73** $\pm$ 1.31	78.13 $\pm$ 2.19	**67.02** $\pm$ 1.94	78.63

The effect of $\mathcal{L}_{\text{RED}}$ on our MIREL As shown in Table 6, we could find that the involvement of RED loss often obtains performance improvements over its counterpart. This is largely because $\mathcal{L}_{\text{RED}}$ can effectively mitigate zero-evidence regions to improve evidential learning, as highlighted in Pandey & Yu (2023).

Table 6. Ablation study on the effect of RED loss on our MIREL (**MNIST-bags**). The base MIL network is ABMIL.

$\mathcal{L}_{\text{RED}}$	Bag-level					Instance-level				
	Acc.	Conf.	OOD-F	OOD-K	$\overline{UE}$	Acc.	Conf.	OOD-F	OOD-K	$\overline{UE}$
×	96.12 $\pm$ 0.37	84.89 $\pm$ 3.49	88.41 $\pm$ 5.15	78.46 $\pm$ 1.76	83.92	85.40 $\pm$ 1.78	96.89 $\pm$ 1.25	63.50 $\pm$ 4.50	57.72 $\pm$ 1.86	72.70
✓	96.48 $\pm$ 0.22	86.63 $\pm$ 1.32	92.84 $\pm$ 0.60	79.95 $\pm$ 4.12	**86.47**	87.71 $\pm$ 0.67	90.73 $\pm$ 1.31	78.13 $\pm$ 2.19	67.02 $\pm$ 1.94	**78.63**

Related UE methods As shown in Table 7, our derived $T(\mathbf{x})$ could often boost the performance of related UE methods in instance-level UE tasks. Moreover, our $T(\mathbf{x})$ surpasses attention-based scoring proxy (a_k) in overall UE performance by 1.64%, 6.41%, and 13.01% for Deep Ensemble, MC Dropout, and $\mathcal{I}$-EDL, respectively. This study further demonstrates the superiority of our $T(\mathbf{x})$ to conventional attention-based instance scoring proxy.

E.3. More Experiments with Different Settings

To investigate the effect of different experimental settings on MIREL's performance, we conduct more experiments and show their results in this section.

Adopting gated attention mechanism in ABMIL When using ABMIL as the base network for our MIREL, we compare two different attention operators proposed in ABMIL, namely, standard attention mechanism and gated attention mechanism. Their results are presented in Table 8. These results show that the standard attention operator is competitive with its gated variant in terms of average UE performance. The standard attention mechanism is our default setting in ABMIL.

Table 7. Additional instance-level ablation study on $T(\mathbf{x})$ for related UE methods (**MNIST-bags**). † These methods directly adopt our $T(\mathbf{x})$ derived from $S(X)$ for instance-level estimation. The other results are copied from Table 1 for comparisons.

Method	Ins.	Instance-level				
		Acc.	Conf.	OOD-F	OOD-K	$\overline{UE}$
Deep Ensemble	a_k	75.56 ± 0.32	71.89 ± 0.91	70.48 ± 0.53	55.22 ± 1.16	65.87
Deep Ensemble †	T	85.97 ± 1.47	84.57 ± 2.62	63.75 ± 2.44	54.22 ± 3.01	**67.51**
MC Dropout	a_k	75.61 ± 0.66	68.40 ± 1.54	68.34 ± 1.06	58.61 ± 1.38	65.12
MC Dropout †	T	88.85 ± 1.54	85.19 ± 3.44	71.61 ± 3.18	57.80 ± 1.53	**71.53**
$\mathcal{I}$-EDL	a_k	75.45 ± 0.13	60.72 ± 1.46	63.91 ± 1.31	54.14 ± 2.19	59.59
$\mathcal{I}$-EDL †	T	85.19 ± 0.64	87.67 ± 1.11	73.52 ± 5.66	56.63 ± 1.66	**72.60**

Table 8. Performance of our MIREL when using standard or gated attention mechanism for ABMIL (**MNIST-bags**).

Attention	Bag-level					Instance-level				
	Acc.	Conf.	OOD-F	OOD-K	$\overline{UE}$	Acc.	Conf.	OOD-F	OOD-K	$\overline{UE}$
Gated	96.52 ± 0.29	87.57 ± 2.51	93.84 ± 2.37	70.67 ± 4.74	84.03	87.96 ± 0.86	87.95 ± 2.46	81.15 ± 2.13	70.45 ± 1.13	**79.85**
Standard	96.48 ± 0.22	86.63 ± 1.32	92.84 ± 0.60	79.95 ± 4.12	**86.47**	87.71 ± 0.67	90.73 ± 1.31	78.13 ± 2.19	67.02 ± 1.94	78.63

Comparison with UE methods on DSMIL The results of this experiment are shown in Table 9. From these results, we observe that our MIREL could still perform better than compared methods in terms of overall UE performance, even when changing the base MIL network from ABMIL to DSMIL. This implies that our method is of good adaptability.

Table 9. Comparison with UE methods when using DSMIL as the base MIL network (**MNIST-bags**). The baseline of this experiment is vanilla DSMIL without any additional UE techniques. Bayes-MIL is not compared here because it is not compatible with DSMIL.

Method	Bag-level					Instance-level				
	Acc.	Conf.	OOD-F	OOD-K	$\overline{UE}$	Acc.	Conf.	OOD-F	OOD-K	$\overline{UE}$
Baseline	96.22 ± 0.17	87.56 ± 0.95	71.13 ± 5.20	60.71 ± 7.91	73.13	70.16 ± 3.56	64.64 ± 0.49	59.75 ± 2.35	57.50 ± 2.55	60.63
Deep Ensemble	96.66 ± 0.17	87.15 ± 0.99	76.06 ± 2.12	64.94 ± 1.49	76.05	72.68 ± 0.84	70.18 ± 0.64	70.15 ± 2.27	64.01 ± 1.65	68.11
MC Dropout	96.36 ± 0.43	88.13 ± 0.61	77.82 ± 2.85	66.56 ± 6.59	77.50	70.27 ± 3.01	64.78 ± 1.33	64.88 ± 5.38	60.78 ± 4.81	63.48
$\mathcal{I}$-EDL	96.60 ± 0.44	89.53 ± 2.03	79.69 ± 9.72	57.77 ± 6.01	75.67	69.04 ± 2.84	63.68 ± 1.43	62.08 ± 2.35	57.93 ± 2.62	61.23
MIREL	96.50 ± 0.37	87.26 ± 2.66	87.27 ± 4.27	62.03 ± 7.78	**78.85**	97.19 ± 0.29	73.79 ± 15.68	73.29 ± 10.85	57.58 ± 3.44	**68.22**

F. Results on CIFAR10-bags

F.1. Main Results

The main comparative results on CIFAR10-bags are shown in Table 10. **(1) Classical deep MIL networks:** Our MIREL could often assist them to perform better in UE. Especially for Max, DSMIL, and ABMIL, the improvements in overall UE performance are 10.45%, 5.11%, and 9.75% at bag level, and 2.80%, 12.06%, and 20.85% at instance level, respectively. **(2) Related UE methods:** With the same base MIL network (ABMIL), our MIREL could often obtain better UE performance than others. Especially at instance level, there is an improvement of 2.40% over the runner-up method in overall UE performance. Moreover, It is worth mentioning that, our MIREL only requires a single forward pass for UE, different from the compared Deep Ensemble, MC Dropout, and Bayes-MIL involving multiple forward passes.

F.2. Uncertainty Analysis

Using ABMIL as the base MIL network, here we show the results of uncertainty analysis on CIFAR10-bags, including bag-level uncertainty (Fig. 11), instance-level uncertainty (Fig. 12), and α_0 distribution (Fig. 13). Our main findings are briefly summarized as follows. (1) The ABMIL models with our MIREL performs slightly better than vanilla ABMIL, in the predictive confidence of abnormal bags. (2) Our MIREL improves the UE capability of ABMIL at instance level by estimating less confidence for OOD instances and more confidence for ID ones. (3) The uncertainty measure of α_0 seems better in detecting the bags with different OOD instance ratios, than that of negative expected entropy.

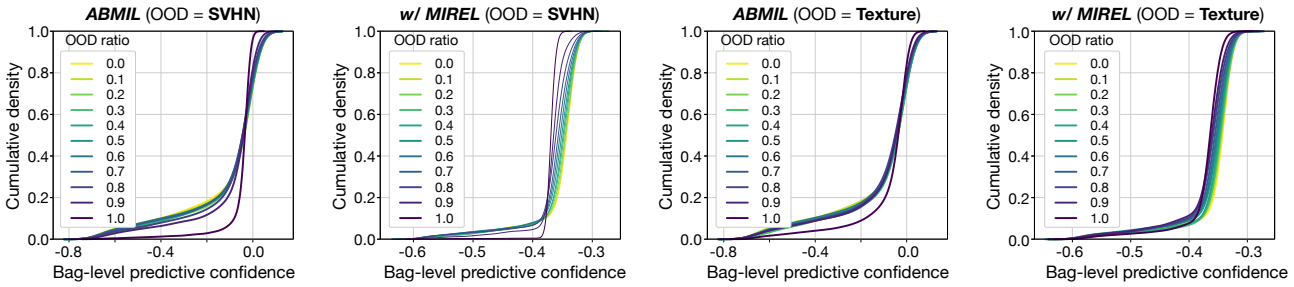

Figure 11. Distribution of bag-level predictive confidence (negative expected entropy). ID dataset is **CIFAR10-bags**.

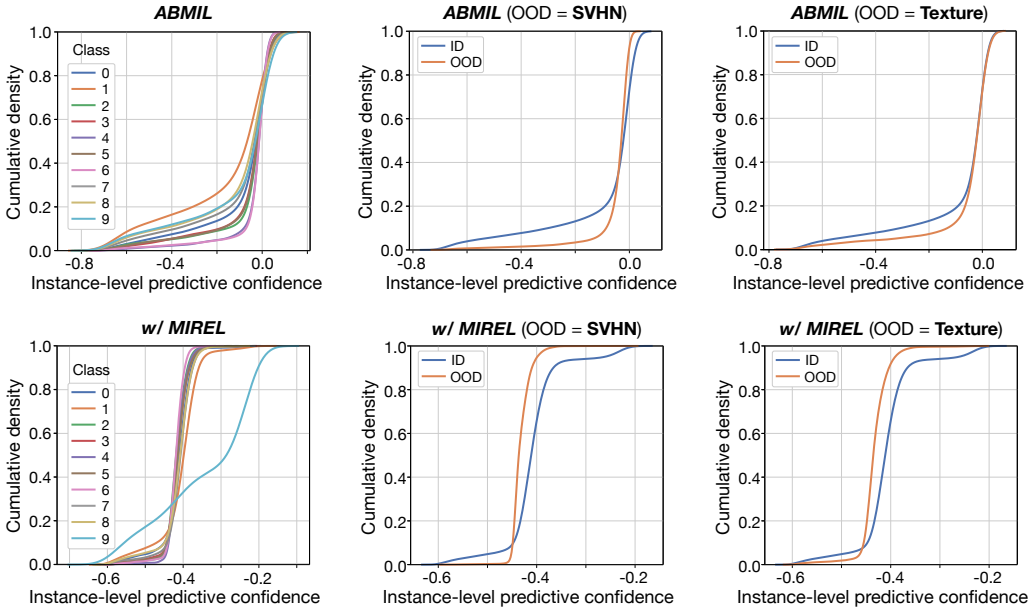

Figure 12. Distribution of instance-level predictive confidence (negative expected entropy). **CIFAR10-bags** is ID dataset.

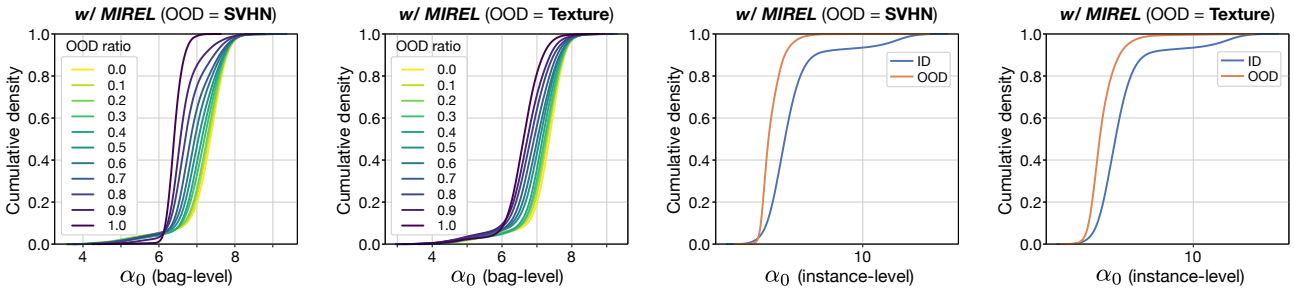

Figure 13. Distribution of bag-level and instance-level α_0 output by the ABMIL models with our MIREL. ID dataset is **CIFAR10-bags**.

Table 10. Main results on **CIFAR10-bags**. OOD-S and OOD-T mean that SVHN and Texture are used for generating OOD bags, respectively. The results colored in gray are from our derived instance estimator $T(\mathbf{x})$. $\overline{\boldsymbol{UE}}$ is the average metrics on three UE tasks.

Method	Bag-level					Instance-level				
	Acc.	Conf.	OOD-S	OOD-T	$\overline{UE}$	Acc.	Conf.	OOD-S	OOD-T	$\overline{UE}$
- Combined with deep MIL networks										
Mean	91.69 ± 0.65	84.79 ± 1.71	87.99 ± 3.09	73.15 ± 4.32	**81.98**	91.92 ± 0.58	56.62 ± 3.32	88.03 ± 2.46	69.10 ± 3.42	**71.25**
Mean + MIREL	92.07 ± 0.80	83.07 ± 1.64	81.05 ± 7.82	73.38 ± 6.47	79.17	92.09 ± 0.75	53.70 ± 11.02	67.06 ± 10.75	63.01 ± 5.68	61.26
Max	92.17 ± 0.52	84.36 ± 0.74	72.39 ± 6.95	65.87 ± 3.95	74.21	90.22 ± 0.68	69.13 ± 1.14	76.65 ± 5.36	67.86 ± 2.70	71.21
Max + MIREL	93.21 ± 0.60	84.96 ± 2.03	90.04 ± 1.40	78.99 ± 3.81	**84.66**	92.81 ± 0.56	68.32 ± 4.98	82.26 ± 5.71	71.44 ± 2.84	**74.01**
DSMIL	92.15 ± 0.85	84.70 ± 0.55	61.81 ± 4.57	62.89 ± 5.04	69.80	71.42 ± 2.24	58.22 ± 0.88	59.10 ± 3.19	53.84 ± 3.42	57.05
DSMIL + MIREL	92.69 ± 0.44	83.02 ± 3.83	77.54 ± 13.29	64.16 ± 7.84	**74.91**	92.71 ± 0.50	60.81 ± 0.55	78.21 ± 8.07	68.30 ± 4.57	**69.11**
ABMIL	91.62 ± 0.62	86.15 ± 1.15	65.56 ± 10.45	63.43 ± 2.84	71.71	76.89 ± 1.07	60.33 ± 0.77	51.60 ± 1.29	44.65 ± 2.43	52.19
ABMIL + MIREL	92.47 ± 0.19	78.43 ± 3.57	88.72 ± 2.78	77.22 ± 6.68	**81.46**	93.18 ± 0.32	66.40 ± 2.48	80.22 ± 4.93	72.49 ± 5.52	**73.04**
- Compared with related UE methods using ABMIL as the base MIL network										
Deep Ensemble	93.33 ± 0.29	86.37 ± 0.91	65.00 ± 6.83	64.86 ± 4.69	72.08	78.80 ± 1.20	71.97 ± 0.64	54.65 ± 4.55	46.07 ± 2.11	57.57
MC Dropout	92.37 ± 0.42	86.26 ± 1.53	62.36 ± 6.62	63.77 ± 3.31	70.79	81.91 ± 1.57	75.81 ± 2.09	64.91 ± 8.65	49.61 ± 3.05	63.44
$\mathcal{I}$-EDL	92.47 ± 0.19	78.43 ± 3.57	88.72 ± 2.78	77.22 ± 6.68	81.46	77.82 ± 0.78	51.61 ± 1.19	62.79 ± 4.12	54.48 ± 2.61	56.29
Bayes-MIL	92.46 ± 0.77	84.52 ± 1.57	83.74 ± 3.82	71.19 ± 5.28	79.82	65.14 ± 32.68	72.44 ± 13.36	74.81 ± 8.78	64.67 ± 6.34	70.64
MIREL	92.47 ± 0.19	78.43 ± 3.57	88.72 ± 2.78	77.22 ± 6.68	**81.46**	93.18 ± 0.32	66.40 ± 2.48	80.22 ± 4.93	72.49 ± 5.52	**73.04**

F.3. Ablation Study

ABMIL with our MIREL The result of this experiment is exhibited in Table 11.

- **Result analysis:** (1) EDL improves the UE capability of vanilla ABMIL models by a large margin (9.75%) at bag level. (2) adopting our derived $T(\mathbf{x})$ rather than a_k for instance prediction often leads to large improvements in overall UE performance, 16.40% and 16.67% for the ABMIL without and with EDL, respectively. (3) our residual instance estimator $R(\mathbf{x})$ shows comparable UE performance with $T(\mathbf{x})$ on CIFAR10-bags.
- **Explanation for the same bag-level performance:** Note that our MIREL obtains the same bag-level performance as its counterparts, *i.e.*, the EDL-based ABMIL network without our $R(\mathbf{x})$. It is because we **only optimize the parameter** π, instead of optimizing both π and ψ, in $\mathcal{L}_{\text{MIREL}}$. We choose to do so as we empirically find that a **deeper instance encoder**, *e.g.*, the network with more than 4 convolutional layers, often leads to unstable training in the weakly-supervised instance-level estimator. One possible reason is that the *weak supervision signals* used for training the instance-level estimator are more likely to vanish in its gradient back-propagation to the deeper layers of instance encoder. Such behavior is also discussed and highlighted in Li et al. (2023). We leave its investigation as future work.
- **Clarification on the setting of instance encoder:** In fact, a deep instance encoder is not a common choice in most MIL applications; instead, a high-dimensional single instance is usually first transformed into a low-dimensional vector and then **a shallow network**, *e.g.* shallow MLP, is utilized as the *real* instance encoder for MIL. This fact can be seen from many real-world MIL applications (Lu et al., 2021; Liu et al., 2024b; Tian et al., 2021; Sapkota & Yu, 2022; Rizve et al., 2023). This means that, in most cases, a shallow instance encoder is a universal setting so our $\mathcal{L}_{\text{MIREL}}$ can be leveraged as expected to optimize both π and ψ and enhance both instance-level and bag-level UE performance.

Table 11. Ablation study on the ABMIL with our MIREL. **CIFAR10-bags** is ID dataset.

Loss		Ins.	Bag-level					Instance-level				
$\mathcal{L}_{\mathcal{I}\text{-EDL}}$	$\mathcal{L}_{\text{MIREL}}$		Acc.	Conf.	OOD-S	OOD-T	$\overline{UE}$	Acc.	Conf.	OOD-S	OOD-T	$\overline{UE}$
		a_k	91.62 ± 0.62	86.15 ± 1.15	65.56 ± 10.45	63.43 ± 2.84	71.71	76.89 ± 1.07	60.33 ± 0.77	51.60 ± 1.29	44.65 ± 2.43	52.19
		T	91.62 ± 0.62	86.15 ± 1.15	65.56 ± 10.45	63.43 ± 2.84	71.71	91.76 ± 0.40	80.54 ± 2.28	67.79 ± 7.60	57.45 ± 5.24	68.59
✓		a_k	92.47 ± 0.19	78.43 ± 3.57	88.72 ± 2.78	77.22 ± 6.68	81.46	77.82 ± 0.78	51.61 ± 1.19	62.79 ± 4.12	54.48 ± 2.61	56.29
✓		T	92.47 ± 0.19	78.43 ± 3.57	88.72 ± 2.78	77.22 ± 6.68	81.46	92.54 ± 0.40	63.83 ± 3.65	82.25 ± 10.73	72.81 ± 8.04	72.96
✓	✓	R	92.47 ± 0.19	78.43 ± 3.57	88.72 ± 2.78	77.22 ± 6.68	**81.46**	93.18 ± 0.32	66.40 ± 2.48	80.22 ± 4.93	72.49 ± 5.52	**73.04**

Optimization strategy for $R(\mathbf{x})$ Similar to that done on MNIST-bags, we test different optimization strategies on CIFAR10-bags. Test results are shown in Table 12. Note that, bag-level results are dropped, since only π is involved in the training of

$R(\mathbf{x})$ (as explained above) and different strategies lead to the same bag-level performance. From Table 12, we could find that $\mathcal{L}_{\text{ins}}^{+}$ often obtains the best UE performance at instance level, surpassing the second-placed $\mathcal{L}_2$ by 1.86% on average. This could further confirm the superiority of our weakly-supervised evidential learning strategy.

Table 12. Ablation study on the loss function used for training $R(\mathbf{x})$. The base MIL network is ABMIL and it is trained on **CIFAR10-bags**.

Loss	Instance-level				
	Acc.	Conf.	OOD-S	OOD-T	$\overline{UE}$
$\mathcal{L}_1$	93.18 ± 0.32	61.26 ± 2.65	75.80 ± 4.89	69.31 ± 5.69	68.79
$\mathcal{L}_2$	93.19 ± 0.34	64.22 ± 3.25	78.42 ± 3.53	70.91 ± 5.28	71.18
$\mathcal{L}_{\text{ins}}^{+}$	93.18 ± 0.32	**66.40** ± 2.48	**80.22** ± 4.93	**72.49** ± 5.52	**73.04**

The effect of $\mathcal{L}_{\text{RED}}$ on our MIREL The results of this experiment are shown in Table 13. From these results, we observe that on CIFAR10-bags, using $\mathcal{L}_{\text{RED}}$ often leads to worse performances in UE, although it is better in overall bag-level UE performance. Nevertheless, we choose to use $\mathcal{L}_{\text{RED}}$ in our baseline approach by default for simplicity. In other words, the setting of $\mathcal{L}_{\text{RED}}$ is simply shared between all experiments and is not fine-tuned for different datasets, although fine-tuning it could lead to better performances in MIUE.

Table 13. Ablation study on the effect of RED loss on our MIREL (**CIFAR10-bags**). The base MIL network is ABMIL.

$\mathcal{L}_{\text{RED}}$	Bag-level					Instance-level				
	Acc.	Conf.	OOD-S	OOD-T	$\overline{UE}$	Acc.	Conf.	OOD-S	OOD-T	$\overline{UE}$
×	92.80 ± 0.41	84.23 ± 2.50	69.05 ± 23.26	78.27 ± 6.84	77.18	93.19 ± 0.38	73.10 ± 2.80	81.57 ± 5.80	74.56 ± 3.23	**76.41**
✓	92.47 ± 0.19	78.43 ± 3.57	88.72 ± 2.78	77.22 ± 6.68	**81.46**	93.18 ± 0.32	66.40 ± 2.48	80.22 ± 4.93	72.49 ± 5.52	73.04

Related UE methods As shown in Table 14, there are large improvements in overall UE performance for compared UE methods, when turning to adopt our $T(\mathbf{x})$ derived from $S(X)$ as instance-level estimator. These improvements are 11.22%, 3.00%, and 16.67% for Deep Ensemble, MC Dropout, and $\mathcal{I}$-EDL, respectively. These again demonstrate our argument, *i.e.*, attention-dependent scoring proxies may not be suitable for instance-level prediction.

Table 14. Additional instance-level ablation study on $T(\mathbf{x})$ for related UE methods (**CIFAR10-bags**). † These methods directly adopt our $T(\mathbf{x})$ derived from $S(X)$ for instance-level estimation. Other results are copied from Table 10 for comparisons.

Method	Ins.	Instance-level				
		Acc.	Conf.	OOD-S	OOD-T	$\overline{UE}$
Deep Ensemble	a_k	78.80 ± 1.20	71.97 ± 0.64	54.65 ± 4.55	46.07 ± 2.11	57.57
Deep Ensemble †	T	93.28 ± 0.16	85.57 ± 1.14	64.92 ± 6.61	55.88 ± 4.33	**68.79**
MC Dropout	a_k	81.91 ± 1.57	75.81 ± 2.09	64.91 ± 8.65	49.61 ± 3.05	63.44
MC Dropout †	T	92.62 ± 0.73	82.65 ± 1.00	62.64 ± 5.92	54.03 ± 3.10	**66.44**
$\mathcal{I}$-EDL	a_k	77.82 ± 0.78	51.61 ± 1.19	62.79 ± 4.12	54.48 ± 2.61	56.29
$\mathcal{I}$-EDL †	T	92.54 ± 0.40	63.83 ± 3.65	82.25 ± 10.73	72.81 ± 8.04	**72.96**

F.4. More Experiments with Different Settings

Similar to those experiments shown in Section E.3, in this section we conduct more experiments with different settings to investigate the effect of these settings on our MIREL scheme.

Gated attention mechanism for ABMIL As shown in Table 15, there is no significant difference in average UE performance between the two attention mechanisms. We choose the standard attention operator by default for ABMIL network in all experiments, because it is more efficient in computation and is adopted more frequently than its gated version although it sometimes performs slightly worse than its gated version in UE tasks.

Comparison with UE methods on DSMIL As shown in Table 16, our MIREL also could often perform better than other UE methods by a large margin at instance level on DSMIL. This result further suggests the adaptability of our method.

Table 15. Performance of our MIREL when using standard or gated attention mechanism for ABMIL (**CIFAR10-bags**).

Attention	Bag-level					Instance-level				
	Acc.	Conf.	OOD-S	OOD-T	$\overline{UE}$	Acc.	Conf.	OOD-S	OOD-T	$\overline{UE}$
Gated	92.25 ± 0.64	81.76 ± 1.42	88.73 ± 4.26	76.54 ± 6.56	**82.35**	93.08 ± 0.17	62.96 ± 3.76	85.98 ± 1.14	73.10 ± 3.23	**74.01**
Standard	92.47 ± 0.19	78.43 ± 3.57	88.72 ± 2.78	77.22 ± 6.68	81.46	93.18 ± 0.32	66.40 ± 2.48	80.22 ± 4.93	72.49 ± 5.52	73.04

Table 16. Comparison with UE methods when using DSMIL as the base MIL network (**CIFAR10-bags**). The baseline of this experiment is vanilla DSMIL without any additional UE techniques. Bayes-MIL is not compared here because it is not compatible with DSMIL.

Method	Bag-level					Instance-level				
	Acc.	Conf.	OOD-S	OOD-T	$\overline{UE}$	Acc.	Conf.	OOD-S	OOD-T	$\overline{UE}$
Baseline	92.15 ± 0.85	84.70 ± 0.55	61.81 ± 4.57	62.89 ± 5.04	69.80	71.42 ± 2.24	58.22 ± 0.88	59.10 ± 3.19	53.84 ± 3.42	57.05
Deep Ensemble	93.20 ± 0.14	86.45 ± 0.74	70.50 ± 4.34	63.57 ± 3.52	73.51	74.25 ± 1.30	66.49 ± 1.01	63.78 ± 4.72	57.86 ± 4.81	62.71
MC Dropout	92.39 ± 0.59	84.81 ± 1.56	72.61 ± 8.90	67.15 ± 6.69	74.86	73.82 ± 1.80	63.18 ± 1.53	66.93 ± 10.46	53.46 ± 6.01	61.19
$\mathcal{I}$-EDL	92.69 ± 0.44	83.02 ± 3.83	77.54 ± 13.29	64.16 ± 7.84	74.91	69.07 ± 8.63	57.98 ± 0.62	52.22 ± 4.02	49.66 ± 4.51	53.29
MIREL	92.69 ± 0.44	83.02 ± 3.83	77.54 ± 13.29	64.16 ± 7.84	**74.91**	92.71 ± 0.50	60.81 ± 0.55	78.21 ± 8.07	68.30 ± 4.57	**69.11**

G. Additional Results on Histopathology Dataset

G.1. Related UE Methods

As shown in Table 17, our $T(\mathbf{x})$ obtains considerable improvements over a_k in overall UE performance, even better than MIREL at instance level. These improvements are 40.08%, 38.57%, and 40.91% for Deep Ensemble, MC Dropout, and $\mathcal{I}$-EDL, respectively. Such impressive results may result from two main factors:

- a_k is obtained by $\mathrm{softmax}$, so its values for negative instances would be extremely small when instance number is very large (recall that there are 11,753 instances in each CAMELYON16 bag on average), thus more likely to yield overconfident estimations.
- our $T(\mathbf{x})$ is directly deduced from $S(X)$, with the ability of distinguishing between negative and positive instances when $S(X)$ is good enough at classifying bags, as stated in Section 4.2.

Table 17. Additional instance-level results of related UE methods on **CAMELYON16**. † These methods directly adopt our $T(\mathbf{x})$ derived from $S(X)$ for instance-level estimation. The other results are copied from Table 3. Particularly, the AUROC of ID instance classification, along with Acc., is reported due to the dominant negative patches (∼ 97.7%) in the slides of CAMELYON16.

Method	Ins.	Instance-level				
		Acc.	AUROC	Conf.	OOD-PRAD	$\overline{UE}$
Deep Ensemble	a_k	96.08 ± 0.02	50.34 ± 2.53	49.62 ± 2.53	28.16 ± 1.07	38.89
Deep Ensemble †	T	89.38 ± 5.23	95.09 ± 0.17	86.36 ± 3.31	71.58 ± 7.25	**78.97**
MC Dropout	a_k	96.05 ± 0.00	56.25 ± 2.16	56.35 ± 2.20	33.93 ± 2.05	45.14
MC Dropout †	T	94.06 ± 2.08	94.07 ± 0.37	88.83 ± 0.85	78.60 ± 3.05	**83.71**
$\mathcal{I}$-EDL	a_k	96.05 ± 0.01	45.39 ± 4.78	45.41 ± 5.33	32.06 ± 1.49	38.74
$\mathcal{I}$-EDL †	T	87.53 ± 5.61	95.11 ± 0.26	87.28 ± 4.12	72.02 ± 7.87	**79.65**

G.2. Distribution Shift Detection

The numerical results of Fig. 3 are presented in Table 18. Similar to that provided in Section 6.3, our result analysis of Table 18 is as follows. (1) **Bag-level**. The AUROC performance on lighter DS is often less than 0.52, namely, lighter DS is hard to detect for all presented UE methods. Our MIREL can detect light DS with an AUROC of 0.59 and strong DS with an AUROC of 0.75, obtaining the best or the second best performance. Moreover, its overall UE performance is the best (0.62). (2) **instance-level**. All compared methods not specially for MIL, consistently obtain an AUROC less than 0.5, indicating meaningless detection results on three shift datasets. On strong DS, our MIREL obtains an AUROC of 0.62, exceeding Bayes-MIL by 5.09%. This experiment could further verify the superiority of our MIREL scheme in MIUE.

Table 18. Comparison with related UE methods on histopathology dataset (**CAMELYON16**). Three shifted versions of CAMELYON16 test set are used for detection. **DS** means Distribution Shift, and '**lighter**', '**light**', and '**strong**' indicate three degrees of shift. The baseline of this experiment is vanilla ABMIL without any additional UE techniques. $\overline{UE}$ is the average metrics on three DS detection tasks.

Method	Bag-level				Instance-level			
	DS-lighter	DS-light	DS-strong	$\overline{UE}$	DS-lighter	DS-light	DS-strong	$\overline{UE}$
Baseline	50.86 ± 0.38	51.87 ± 1.36	54.48 ± 6.31	52.40	49.17 ± 0.35	46.87 ± 0.53	42.88 ± 1.87	46.31
Deep Ensemble	50.70 ± 0.40	52.39 ± 1.23	53.62 ± 9.06	52.24	49.09 ± 0.18	45.86 ± 0.87	40.73 ± 0.94	45.23
MC Dropout	50.49 ± 0.49	52.54 ± 1.42	52.70 ± 5.35	51.91	49.68 ± 0.61	47.80 ± 1.23	43.89 ± 2.07	47.12
$\mathcal{I}$-EDL	51.50 ± 0.34	57.46 ± 2.48	72.74 ± 3.58	60.57	49.11 ± 0.33	45.84 ± 1.51	38.89 ± 2.84	44.61
Bayes-MIL	50.19 ± 0.46	52.20 ± 1.45	**77.42** ± 6.32	59.94	**50.29** ± 0.21	50.36 ± 1.19	56.95 ± 3.92	52.53
MIREL	**51.98** ± 0.89	**58.97** ± 2.14	74.84 ± 1.71	**61.93**	50.11 ± 0.18	**50.72** ± 0.75	**62.04** ± 1.19	**54.29**

H. Synthetic MIUE Experiment

To understand the UE behavior of our *weakly-supervised* instance estimator, we synthesize a simple bag dataset with 2-dimensional instances. The surface of predictive probability and predictive uncertainty are visualized in a 2D plane for intuitive interpretation.

H.1. MIL Dataset

2D instance generation We generate 2D instances using three isotropic Gaussian with $\sigma^2 = 0.1$. The centroid of three Gaussian are located in (0, 1.5), (–1.5, –0.5), and (1.5, –0.5). Each Gaussian contains 1,000 points (instances). The Gaussian with a centroid of (0, 1.5) is the class of interest (positive), and the remaining two are negative. Note that, actually, instance labels are *unknown* in training.

Bag generation Following the process described in Appendix D.1, we synthesize bags using the 2D instances generated above. Finally, there are 2,000 bags for training and 500 bags for validation and early stopping. Only bag-level labels are utilized for training MIL networks.

H.2. Implementation Details

ABMIL is adopted as the base MIL network in this experiment. Its instance encoder is implemented by an MLP with two layers. In network training, learning rate is set to 5×10^{-5} and $\lambda_1 = 0.4$. The patience step for learning rate decay and early stopping are 5 and 10 epochs, respectively. Other settings are the same as those given in Appendix D.3.

H.3. Result Visualization

Similar to the settings of ablation study, there are three models used for analysis and comparison, as explained in Table 4. Their results are visualized in Fig. 14. Our main observations are as follows:

- For the a_k of ABMIL, it shows high confidence in the region near and below negative instances, but low in the region near or above positive ones. It is mainly caused by the $\mathrm{softmax}$ operator in attention score calculation. Generally, $\mathrm{softmax}$ would lead to small a_k for positive instances when multiple positive instances are contained in a bag. As a result, there would be relatively large entropy and high uncertainty for positive instances, compared to negative ones.

- For the $T(\mathbf{x})$ of ABMIL, it seems better than a_k in instance classification. However, it only predicts high uncertainty near the boundary between positive and negative instances, showing overconfidence in the region far from ID instances. This behavior is very similar to that of standard classification models, possibly caused by the ignorance of epistemic uncertainty or distributional uncertainty in predictive modeling.

- For the $T(\mathbf{x})$ of EDL-based ABMIL, it captures the uncertainty in some regions far from ID instances, owing to its Dirichlet-based predictive uncertainty modeling. For Dirichlet-based models, the uncertainty caused by the distributional mismatch between training and test is specially considered and incorporated into model prediction (Malinin & Gales, 2018; Ulmer et al., 2023).

- For the $R(\mathbf{x})$ of MIREL-based ABMIL, it further improve the quality of predictive uncertainty. Especially in the region near positive instances, $R(\mathbf{x})$ often predicts less uncertainty than the $T(\mathbf{x})$ of EDL-based ABMIL. This improvement is largely due to our residual evidential learning scheme. As stated in Section 4.3, our $R(\mathbf{x})$ is specially proposed to learn instance-specific residuals and is encouraged to compensate for the initial biased $T(\mathbf{x})$.

✍ **Discussion** This synthetic MIL experiment could assist us in understanding the UE behavior of different weakly-supervised instance estimators. At the same time, it could be found that there is still room for further improvements. For example, $R(\mathbf{x})$ cannot obtain desirable UE results in some regions far from negative instances. This could be one of the main challenges posed by weak supervision. We leave its solution as future work.

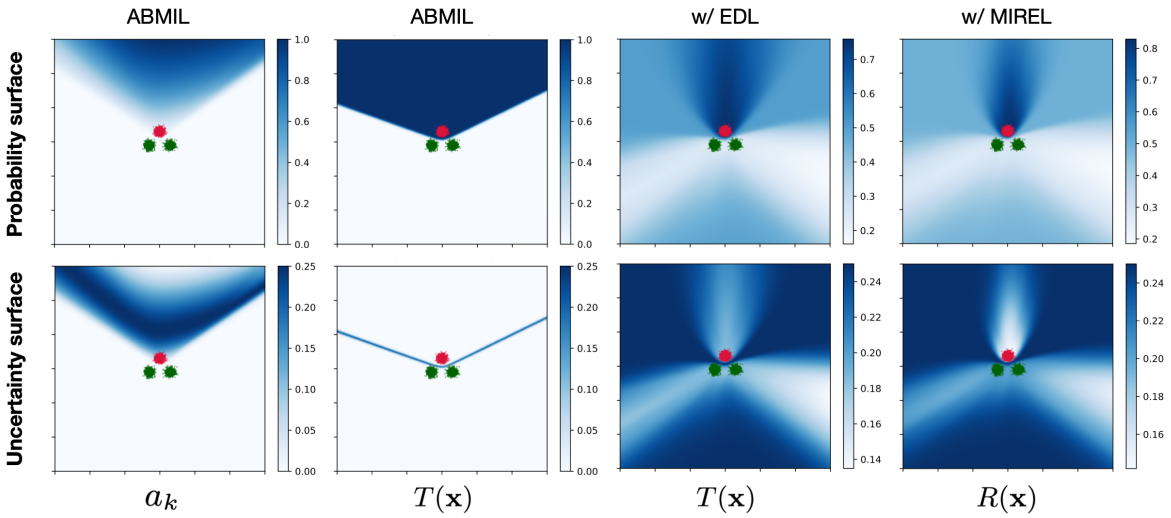

Figure 14. Visualization of the instance prediction given by different weakly-supervised estimators. A synthetic MIL dataset with 2D instances is used in this experiment. The points colored in red and green are positive and negative instance, respectively. Note that, unlike standard fully-supervised settings, there are no complete instance labels to use for training.