CONFORMITY SUPERVISED EVIDENTIAL DEEP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

While deep neural networks achieve remarkable performance, their tendency to produce overconfident predictions remains a critical issue. Evidential deep learning (EDL) solves this problem by modeling class probabilities with a Dirichlet distribution and quantifying the uncertainty in the model's predictions. However, conventional EDL lacks explicit supervision for uncertainty quantification, which leads to a risk of assigning excessive evidence to unreliable samples. In this paper, we propose Conformity Supervised Evidential Deep Learning (CSEDL), an EDL framework that learns class prediction and uncertainty through complementary supervision. Specifically, we introduce an extra uncertainty head and the rank of nonconformity score to supervise it without additional annotations. Together with the direction head supervised by ground-truth labels, CSEDL forms a dual-head architecture, and the two heads are then combined to reconstruct Dirichlet parameters for final prediction. Extensive experiments on out-of-distribution (OOD) detection, tumor referral prediction, autonomous-driving trajectory risk forecasting, trusted multi-view classification and calibrating large language models (LLMs) indicate that CSEDL improves uncertainty quantification ability significantly across multiple tasks. Code is available in the supplementary materials.

1 INTRODUCTION

Deep neural networks have achieved significant performance across a wide range of tasks (LeCun et al., 2015), but their tendency to provide overconfident predictions remains a challenge (Guo et al., 2017). To address this challenge, a machine learning model must go beyond accurate prediction and identify when its predictions are reliable (Guo et al., 2017; Jiang et al., 2018), especially in safety crucial applications such as medical diagnosis (Leibig et al., 2017; Kompa et al., 2021) and autonomous driving (McAllister et al., 2017).

Evidential Deep Learning (EDL) (Sensoy et al., 2018) provides an appealing framework for this challenge by representing categorical predictions with a Dirichlet distribution to quantify predictive uncertainty. Specifically, EDL constructs an evidential network to generate evidence for each sample and supervises the output class direction of the network with ground-truth labels. Then the uncertainty of each sample is calculated by its total evidence, where larger evidence indicates lower uncertainty.

However, EDL supervises the class direction by the ground-truth label solely without providing an explicit target for generating an appropriate amount of evidence and uncertainty. As the training goes, EDL tends to assign excessive evidence and little uncertainty on samples regardless of whether they are uncertain or not (a detailed derivation is provided in Eq. (3) of Section 3). The left part of Figure 1 illustrates this phenomenon. As the training proceeds in EDL, the evidence of in-distribution (ID, certain) samples and out-of-distribution (OOD, uncertain) samples increases simultaneously. Consequently, excessive evidence on both ID and OOD samples leads to a bad discrimination in uncertainty between them where $u_{\text{OOD}} - u_{\text{ID}} = 0.0148$.

To generate better uncertainty in EDL, for the first time, we propose to supervise it explicitly by introducing an additional uncertainty head in the evidential network. Specifically, a separate target for uncertainty prediction is constructed to prevent evidence from growing excessively while continuing to use ground-truth labels to supervise class prediction. Ideally, if a ground-truth uncertainty u_i^* was available for the i-th training sample, the two quantities could be learned separately: the ground-truth class label y_i would supervise the predictive direction, while u_i^* would supervise the uncertainty.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

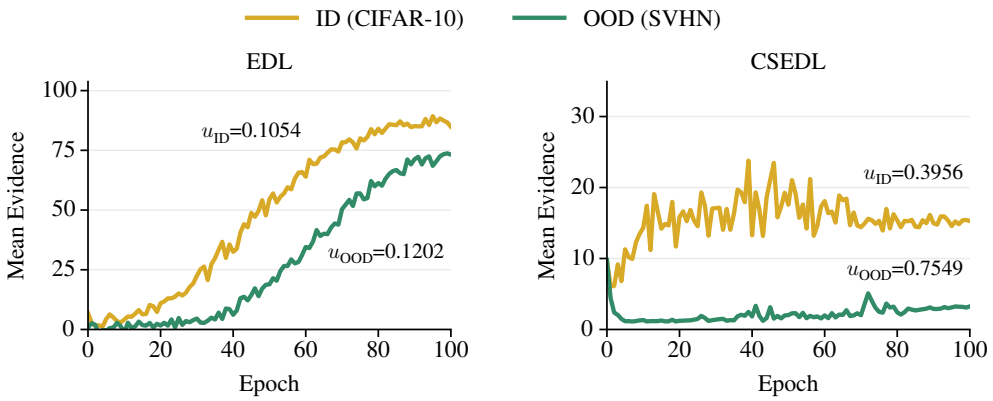


Figure 1: Comparison of EDL and CSEDL in the mean evidence on in-distribution (ID, certain) and out-of-distribution (OOD, uncertain) samples during training. u_{ID} and u_{OOD} represent the mean uncertainty of ID and OOD samples respectively, in which larger evidence indicates lower uncertainty. In conventional EDL (left), evidence on ID and OOD datasets increases simultaneously to a large value, degrading the discrimination of uncertainty between ID and OOD samples. In our proposed method CSEDL (right), evidence on ID samples increases and then becomes stable while OOD samples keep low evidence, showing a clear distinction between uncertainty in ID and OOD samples.

However, such uncertainty annotations are generally unavailable. A training sample is typically observed with only one class label, rather than an accurate value of its uncertainty. Therefore, the KEY QUESTION is how to construct an uncertainty target from ordinary labeled data without requiring additional annotations.

Traditional sample difficulty metrics like maximum softmax probability (Hendrycks & Gimpel, 2017) and predictive entropy (Malinin & Gales, 2018) can serve as an uncertainty target without extra annotations by assigning more uncertainty to more difficult samples. However, these metrics are unrelated to ground-truth and may assign low uncertainty to confidently incorrect predictions. Nonconformity score in conformal prediction (Vovk et al., 2005; Shafer & Vovk, 2008; Barber & Tibshirani, 2026) measures the agreement of the prediction with the ground-truth label on target class and can identify confidently incorrect predictions. As nonconformity scores may be clustered around a certain value, we further select the rank of nonconformity score to provide a smooth and uniform uncertainty target. The effectiveness of this design is further discussed in Section 4.2.

In this work, we propose a new EDL framework named **Conformity Supervised Evidential Deep Learning (CSEDL)**, in which class labels guide class prediction while nonconformity score provides an additional training signal for uncertainty. Additionally, motivated by this complementary supervision structure of class prediction and uncertainty, we also redesign the last layer of the evidential network with two heads. The direction head is supervised by the ground-truth class label, whereas the uncertainty head is supervised by the empirical rank of the nonconformity score, and the two heads are then combined to reconstruct the Dirichlet parameters to obtain final prediction. The right part of Figure 1 shows the effectiveness of CSEDL, where the evidence on ID samples increases and then becomes stable while that on OOD samples keeps low. As a result, a clear distinction between uncertainty in ID and OOD samples forms, where $u_{\text{OOD}} - u_{\text{ID}} = 0.3593$. A flowchart of CSEDL is in Figure 3 of Appendix A.

Our main contributions are summarized as follows:

- We observe the problem that conventional EDL can assign excessive evidence on unreliable samples as the training goes, which leads to inaccurate uncertainty and decreases its performance.
- To solve the problem above, we propose Conformity Supervised Evidential Deep Learning (CSEDL), an EDL model with separate heads for the class prediction and the uncertainty. We further introduce the rank of nonconformity score to supervise the uncertainty head.
- Our proposed method achieves superior performance in extensive experiments across OOD detection, tumor referral prediction, autonomous-driving trajectory risk forecasting, trusted multi-

view classification under missing views, and confidence calibration for large language models (LLMs), outperforming other state-of-the-art methods.

2 Related Works

2.1 Uncertainty Estimation in Deep Learning

Deep neural networks have achieved remarkable performance in various domains, yet they often produce overconfident predictions when encountering ambiguous or out-of-distribution samples. Reliable uncertainty estimation has therefore become an essential component for deploying deep models in safety-critical applications. Existing uncertainty estimation approaches can be broadly categorized into Bayesian methods, ensemble-based methods, and deterministic approaches.

Bayesian neural networks (BNNs) (Blundell et al., 2015) estimate uncertainty by placing distributions over model parameters and approximating the posterior distribution through sampling or variational inference. Deep ensembles (Lakshminarayanan et al., 2017) train multiple independently initialized models and measure the predictive diversity among them. Despite their strong empirical performance, they require multiple networks and consequently increase storage and inference costs.

Several approaches have explored how to achieve uncertainty estimation with a single deterministic model. Posterior Network (PostN) (Charpentier et al., 2020) combines latent-space density estimation using normalizing flows with density-based pseudo-counts to construct sample-dependent posterior distributions, enabling uncertainty estimation without requiring OOD samples during training. Natural Posterior Network (NatPN) (Charpentier et al., 2021) further generalizes this idea to exponential-family target distributions by combining latent density estimation with conjugate Bayesian updates, providing efficient uncertainty estimation for both classification and regression.

2.2 Evidential Deep Learning

Among deterministic distributional approaches, Evidential Deep Learning (EDL) (Sensoy et al., 2018) provides an attractive framework by predicting parameters of a higher-order distribution instead of directly estimating class probabilities. It formulates classification as a subjective evidence accumulation process, where neural networks predict class-wise evidence and construct a Dirichlet distribution over categorical probabilities. By representing uncertainty through the concentration of the Dirichlet distribution, EDL provides an efficient alternative to Bayesian inference and ensemble-based uncertainty estimation.

Following the original formulation, subsequent studies have improved evidential learning from different perspectives, which is detailed in Appendix B.

2.3 Conformal Prediction for Uncertainty Quantification

Conformal Prediction (CP) (Vovk et al., 2005; Shafer & Vovk, 2008; Barber & Tibshirani, 2026) provides a distribution-free framework for quantifying predictive uncertainty by constructing prediction sets or intervals with statistical guarantees. Unlike Bayesian approaches that rely on assumptions about model parameters, CP estimates uncertainty through nonconformity scores derived from model outputs and calibration data. This property makes CP broadly applicable to arbitrary predictive models and has motivated its adoption in various uncertainty-aware learning scenarios.

However, existing CP-based methods typically treat conformal prediction as a post-hoc calibration procedure applied after model training. The conformal score is used to adjust prediction sets rather than to provide a direct training signal for learning predictive uncertainty.

3 Proposed Method

Preliminaries. Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the input space of a K-class classification problem and $\mathcal{Y} = \{1, \ldots, K\}$ the corresponding label space. Given an input $x \in \mathcal{X}$, its ground-truth label is denoted by $y \in \mathcal{Y}$, and the corresponding one-hot vector by $\mathbf{y} \in \{0, 1\}^K$.

Conventional EDL Framework. Conventional EDL represents categorical predictions of sample x with a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha}(\boldsymbol{x})) = \mathrm{Dir}([\alpha_1(x), \alpha_2(x), ..., \alpha_K(x)])$. An evidential network $h_\theta(\cdot)$ directly predicts non-negative class-wise evidence vector $\mathbf{e}(x) = \mathrm{softplus}(h_\theta(x)) \in \mathbb{R}^K_{\geq 0}$ and constructs the Dirichlet parameters as $\boldsymbol{\alpha}(x) = \mathbf{e}(x) + \mathbf{1}$, where θ is the parameters of the evidential network and $\mathbf{1}$ represents an all-one vector. The prediction vector $\mathbf{p}(x)$ and uncertainty $u(x)$ are calculated as $\mathbf{p}(x) = \frac{\boldsymbol{\alpha}(x)}{S(x)}$ and $u(x) = \frac{K}{S(x)}$, where $S(x) = \sum_{k=1}^K \alpha_k(x) = \sum_{k=1}^K e_k(x) + K$ is the sum of Dirichlet parameters, which can also be roughly regarded as the sum of evidence. The loss function of EDL consists of two parts, where $\mathcal{L}_{\mathrm{MSE}}$ supervises the prediction $\mathbf{p}(x)$ with the ground-truth label $\mathbf{y}$:

$$\mathcal{L}_{\mathrm{MSE}} = \|\mathbf{y} - \mathbf{p}(\mathbf{x})\|_2^2 + \frac{1 - \|\mathbf{p}(\mathbf{x})\|_2^2}{S(x) + 1}. \tag{1}$$

and $\mathcal{L}_{\mathrm{KL}}$ suppresses the evidence of non-target classes:

$$\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}[\mathrm{Dir}(\mathbf{p}; \tilde{\boldsymbol{\alpha}}) \parallel \mathrm{Dir}(\mathbf{p}; \mathbf{1})], \quad \text{s.t. } \tilde{\boldsymbol{\alpha}} = \mathbf{y} + (\mathbf{1} - \mathbf{y}) \odot \boldsymbol{\alpha}, \tag{2}$$

The final objective of EDL is defined as $\mathcal{L}_{\mathrm{EDL}} = \mathcal{L}_{\mathrm{MSE}} + \lambda \mathcal{L}_{\mathrm{KL}}$, where λ is a hyperparameter. However in conventional EDL, while keeping the prediction vector $\mathbf{p}(x)$ unchanged,

$$\frac{\partial \mathcal{L}_{\mathrm{MSE}}}{\partial S(x)} = -\frac{1 - \|\mathbf{p}(x)\|_2^2}{(S(x) + 1)^2} < 0, \tag{3}$$

which indicates that L_{MSE} decreases monotonically as the sum of Dirichlet parameters $S(x)$ increases. While $\mathcal{L}_{\mathrm{KL}}$ only suppresses the evidence of non-target classes and leaves the target class uncontrolled, the evidence neural network can be over-evidencing after training, in which it provides very large evidence for any input. As a result, this ill-posed EDL model may assign excessive evidence and little uncertainty on samples regardless of whether they are uncertain or not.

3.1 Dual-Head Parameterization

To solve the problem above in conventional EDL, CSEDL supervises uncertainty explicitly and factorizes class prediction and its uncertainty into two separate output branches. Given the feature representation $h_\theta(x)$, the direction head $\boldsymbol{\pi}_\theta(x)$ predicts a vector

$$\boldsymbol{\pi}_\theta(x) = \mathrm{softmax}(g_\theta(h_\theta(x))), \tag{4}$$

while the uncertainty head $q_\theta(x)$ predicts a scalar logit

$$q_\theta(x) = \sigma(a_\theta(h_\theta(x))) \in (0, 1), \tag{5}$$

where $g_\theta(\cdot)$ maps the feature representation $h_\theta(x)$ to a class logit vector, $a_\theta(\cdot)$ maps $h_\theta(x)$ to a scalar uncertainty and $\sigma(z) = 1/(1 + \exp(-z))$ denotes the sigmoid function. Then $q_\theta(x)$ is mapped to a bounded evidential uncertainty as

$$u_\theta(x) = u_{\min} + (u_{\max} - u_{\min}) q_\theta(x), \qquad 0 < u_{\min} < u_{\max} \leq 1, \tag{6}$$

where we set $u_{\min} = 0.05$ and $u_{\max} = 0.95$ in our implementation. Given $u(x) = \frac{K}{S(x)}$, $S(x)$ is reconstructed as $S(x) = \frac{K}{u_\theta(x)}$. Assuming that the total evidence $\sum_{k=1}^K e_k(x) = S(x) - K$ is assigned to the k-th class with a probability of $\pi_{\theta,k}(x)$, we have

$$\alpha_k(x) = 1 + (S(x) - K) \cdot \pi_{\theta,k}(x). \tag{7}$$

The prediction on the k-th class $p_{\theta,k}(x)$ can be consequently expressed as

$$p_{\theta,k}(x) = \frac{\alpha_{\theta,k}(x)}{S_\theta(x)} = \frac{u_\theta(x)}{K} + \big(1 - u_\theta(x)\big)\pi_{\theta,k}(x). \tag{8}$$

3.2 Direction Supervision

The ground-truth label can provide an explicit supervision signal for the direction head simply through

$$\mathcal{L}_{\text{dir}}(x_i, y_i) = -\log \pi_{\theta, y_i}(x_i), \tag{9}$$

where $\pi_{\theta,y_i}(x_i)$ represents the prediction on the ground-truth label of the i-th sample. However, $\mathcal{L}_{\text{dir}}$ does not provide an explicit sample-wise target specifying how uncertain an individual prediction should be. We therefore introduce another supervision for the uncertainty head.

3.3 Uncertainty Supervision

Besides direction supervision, CSEDL introduces conformal-rank supervision to provide the uncertainty head with a sample-wise uncertainty target.

For the i-th labeled sample (x_i, y_i), we first define its nonconformity score as

$$Z_i = 1 - \text{sg}[p_{\theta, y_i}(x_i)], \tag{10}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator and $p_{\theta,y_i}(x_i)$ denotes the predictive probability assigned to the ground-truth class. Larger nonconformity scores indicate less reliable predictions. Thus, samples with different predictive reliabilities can be distinguished even when they share the same class label.

The absolute magnitude of nonconformity scores can vary with many task and model dependent factors, making them unsuitable to supervise the uncertainty. To this end, we adopt the relative rank of nonconformity scores to reflect how atypical a sample is compared with historical predictions. Let

$$F(z) = \mathbb{P}\big(Z(X, Y) \le z\big) \tag{11}$$

denote the cumulative distribution function of the nonconformity score. For the i-th sample (x_i, y_i), its conformal rank in this distribution is defined as

$$R_i = F(Z(x_i, y_i)). \tag{12}$$

A larger rank indicates that the sample lies in the upper tail of the nonconformity distribution and is therefore relatively difficult for the current predictor. Compared with the raw nonconformity score, the rank provides a smooth and uniform measure of uncertainty and guarantees a better uncertainty quantification performance, preventing the uncertainty of samples from being clustered around a certain value. *The effectiveness of this design is further verified in Table 1 of Section 4.2.*

The conformal rank in Eq. (12) can be obtained by computing the nonconformity scores over the entire training set before every batch in each epoch but it is too time-consuming (about 30 minutes per epoch in our environment). An approximation would be to recompute the nonconformity scores over the entire training set at the beginning of each epoch t and construct an empirical score distribution $\widehat{F}_t$. However, such a full-dataset forward pass also introduces severe computational overhead (see Appendix F). Moreover, as the model parameters are updated across batches within the epoch, $\widehat{F}_t$ gradually becomes stale and may no longer accurately reflect the current nonconformity distribution.

Therefore, we maintain a historical nonconformity score cache $\mathcal{M}$ that stores nonconformity scores from previous batches. $\mathcal{M}$ is updated after every batch as a first-in-first-out queue: newly generated scores are appended after each iteration, while the oldest ones are removed when the cache exceeds its maximum size. Therefore, the empirical conformal rank of the i-th sample is estimated as

$$\widehat{R}_i = \frac{1}{|\mathcal{M}^{(t)}|} \sum_{v \in \mathcal{M}^{(t)}} \sigma\left(\frac{Z_i - v}{T}\right), \tag{13}$$

where the cache size $|\mathcal{M}|$ and the smoothing temperature T are two hyperparameters. The historical cache provides a more stable approximation of the nonconformity distribution than estimating the rank from only the batch data of the current iteration.

Finally, the uncertainty head $q_\theta(x)$ is supervised by this target:

$$\mathcal{L}_{\text{unc}} = \frac{1}{|B|} \sum_{i \in B} \Big(q_\theta(x_i) - \text{sg}[\widehat{R}_i]\Big)^2, \tag{14}$$

where B denotes the sample index set in a batch and $|B|$ denotes its size. Therefore, samples with larger nonconformity ranks encourage the uncertainty head to output larger uncertainty values.

3.4 Final Training Objective

Following conventional EDL, we additionally regularize non-target evidence using

$$\mathcal{L}_{\mathrm{kl}} = D_{\mathrm{KL}}\left(\mathrm{Dir}(\widetilde{\boldsymbol{\alpha}}) \,\|\, \mathrm{Dir}(\mathbf{1})\right), \quad \text{s.t. } \widetilde{\boldsymbol{\alpha}} = \mathbf{y} + (\mathbf{1} - \mathbf{y}) \odot \boldsymbol{\alpha}. \tag{15}$$

The complete training objective is

$$\mathcal{L}_{\mathrm{CSEDL}} = \mathcal{L}_{\mathrm{dir}} + \lambda_{\mathrm{unc}}\mathcal{L}_{\mathrm{unc}} + \lambda_{\mathrm{kl}}(t)\mathcal{L}_{\mathrm{kl}}, \tag{16}$$

where $\mathcal{L}_{\mathrm{dir}}$ provides explicit class-direction supervision, $\mathcal{L}_{\mathrm{unc}}$ provides conformal-rank supervision for the uncertainty head, and $\mathcal{L}_{\mathrm{kl}}$ is the conventional Dirichlet KL regularizer. λ_{unc} is a hyperparameter. The KL coefficient $\lambda_{\mathrm{kl}}(t)$ is annealed during the warm-up stage, where $\lambda_{kl}(t) = \min(1, \frac{t}{10})$. Pseudo-code of CSEDL is provided in Appendix C.

4 Experiments

4.1 Experimental Setup

Baselines. We primarily compare CSEDL with evidential deep learning methods, including conventional EDL, $\mathcal{I}$-EDL (Deng et al., 2023), DAEDL (Yoon & Kim, 2024), HEDL (Qu et al., 2024), $\mathcal{F}$-EDL (Yoon & Kim, 2026), R-EDL (Chen et al., 2024) and Re-EDL (Chen et al., 2025). For out-of-distribution (OOD) detection, we additionally compare with Deep ensembles (DEs, $M = 5$) (Lakshminarayanan et al., 2017), Posterior Network (PostN) (Charpentier et al., 2020) and Natural Posterior Network (NatPN) (Charpentier et al., 2021), which provide competitive uncertainty estimation outside the conventional EDL framework. For LLMs calibration, we further compare with IB-EDL (Li et al., 2025), an EDL method that is specifically designed for calibrating LLMs.

Datasets. For OOD detection, CIFAR-10 (Krizhevsky, 2009) is used as the in-distribution (ID) dataset, while SVHN (Netzer et al., 2011), CIFAR-100 (Krizhevsky, 2009), GTSRB (Stallkamp et al., 2011), LFWPeople (Huang et al., 2007), Places365 (Zhou et al., 2018) and Food101 (Bossard et al., 2014) are the OOD datasets. For tumor referral prediction, we use the PatchCamelyon (Veeling et al., 2018) dataset. For trajectory risk forecasting, we use the Argoverse 1 Motion Forecasting dataset (Chang et al., 2019). For trusted multi-view classification, we use the Handwritten (Duin, 1998) multi-view dataset. For LLM calibration, we choose ARC-C (Clark et al., 2018), ARC-E (Clark et al., 2018), OBQA (Mihaylov et al., 2018) and SciQ (Welbl et al., 2017) as datasets.

The details of the datasets are in Appendix D. The trusted multi-view classification results, a sensitivity analysis of the maximum cache size $M_{\max}$, and an analysis of the evolution of uncertainty distributions during training are presented in Appendices E, F and G respectively due to page limit.

Implementation. In our settings, λ_{unc} is set to 0.01 and the smoothing temperature T is set to 0.05. The maximum cache size is set to $M_{\max} = 1024$, which is explained in Appendix F.

4.2 OOD Detection

We use CIFAR-10 as the ID dataset and evaluate each method against six OOD datasets. To further verify the effectiveness of the conformal rank in our proposed method, we construct two ablation baselines, ***Entropy*** and ***Raw-score***. They replace the conformal rank with the rank of predictive entropy (Malinin & Gales, 2018) and the raw nonconformity score, respectively. We report in-distribution classification accuracy (ID Acc, %) and AUROC on identifying out-of-distribution samples (OOD AUROC, %) and the average OOD AUROC (Avg., %).

As shown in Table 1, CSEDL outperforms all other baselines except OOD AUROC on Places365 across all the cases. Notably, CSEDL shows particularly strong OOD detection performance, improving the AUROC on LFWPeople by 6.34 percentage points over the second-best baseline (DEs). Moreover, supervising uncertainty with the rank of predictive entropy or nonconformity score instead of conformal rank leads to a decline in performance. This is because the nonconformity score is a better target than the entropy, and the raw scores of samples often cluster around some certain values, degrading the discrimination of ID and OOD samples. The conformal rank provides a smooth and uniform measure of uncertainty, and thus improves the performance.

Table 1: OOD detection performance on CIFAR-10 measured by AUROC (↑). Results are reported as mean ± standard deviation except Average. The best results are in **bold**.

Method	ID Acc	SVHN	CIFAR100	GTSRB	LFWPeople	Places365	Food101	Avg.
DEs	92.33±0.63	93.59±0.37	87.33±0.22	88.59±0.39	93.10±0.96	**87.55±0.41**	86.58±0.58	89.46
PostN	87.82±0.42	86.81±0.88	83.18±0.55	85.54±0.79	90.25±0.82	75.59±1.04	80.80±0.82	83.70
NatPN	87.73±0.39	86.60±0.91	82.90±0.58	85.30±0.81	90.40±0.78	74.80±1.09	81.00±0.80	83.50
EDL	86.76±0.54	79.84±1.24	79.20±0.83	76.24±1.36	71.21±1.58	79.35±0.72	78.56±0.91	77.40
$\mathcal{I}$-EDL	90.40±1.23	92.09±0.47	84.16±0.91	85.00±1.02	83.44±0.66	83.42±1.37	83.81±0.58	85.32
DAEDL	89.15±0.84	91.90±1.15	83.84±0.52	89.29±1.43	78.14±0.69	82.77±0.95	83.87±1.28	84.97
HEDL	88.76±1.08	85.45±0.73	82.82±1.56	87.30±0.44	68.94±0.87	87.07±1.19	84.80±0.52	82.73
$\mathcal{F}$-EDL	89.01±0.67	90.78±1.31	85.02±0.48	87.26±0.83	79.05±1.74	85.56±0.59	82.23±1.02	84.98
R-EDL	86.50±0.47	89.24±0.75	82.30±0.66	86.42±0.72	91.80±0.69	82.56±0.67	80.79±0.73	85.52
Re-EDL	86.23±0.45	90.49±0.68	83.24±0.62	86.38±0.70	90.90±0.74	83.87±0.63	82.77±0.65	86.28
Entropy	90.41±0.66	89.50±0.41	80.11±0.93	77.07±0.58	82.35±0.82	78.36±0.89	79.94±0.46	81.22
Raw-score	88.40±0.45	66.51±0.45	57.84±0.68	53.42±0.62	66.38±0.70	76.83±0.74	69.97±0.63	65.16
CSEDL	**92.50±0.32**	**97.01±0.41**	**87.59±0.52**	**90.68±0.58**	**99.44±0.18**	86.81±0.48	**87.06±0.61**	**91.43**

Table 2: Tumor referral prediction performance on PCam. Results are reported as mean ± standard deviation. The best results are in **bold**.

Method	Acc↑	DiagAUROC↑	ErrAUROC↑	Missed Tumors Captured↑		Post-Referral Acc↑	
				@5%	@10%	@5%	@10%
EDL	78.85±0.41	86.43±0.36	65.83±0.82	8.33±0.69	19.12±1.02	79.24±0.38	82.30±0.44
$\mathcal{I}$-EDL	81.83±1.24	91.05±0.67	72.88±1.05	11.73±0.36	23.57±0.72	84.22±1.51	86.44±0.93
DAEDL	81.58±1.12	91.31±0.74	57.11±1.36	1.05±0.09	2.36±0.15	82.02±0.85	82.45±1.06
HEDL	80.98±1.32	88.43±0.58	68.61±1.04	3.74±0.22	9.48±0.38	82.76±1.20	84.58±0.69
$\mathcal{F}$-EDL	80.82±1.08	92.20±0.63	74.36±0.92	10.30±0.37	20.44±0.65	82.98±1.47	85.03±0.88
R-EDL	82.09±0.39	90.31±0.34	72.08±0.71	11.22±0.72	22.34±0.93	84.40±0.36	86.51±0.41
Re-EDL	81.39±0.42	88.15±0.40	67.94±0.88	11.74±0.66	22.14±0.98	83.78±0.39	85.92±0.45
CSEDL	**84.59±0.31**	**92.57±0.29**	**75.15±0.63**	**13.53±0.58**	**26.80±0.79**	**86.84±0.32**	**89.04±0.35**

4.3 Tumor Referral Prediction on PatchCamelyon

We further evaluate CSEDL in a medical image analysis setting using PatchCamelyon (PCam) (Veeling et al., 2018), in which each sample is a small tissue patch, and the label indicates whether tumor tissues are present in the patch. This setting is built to simulate a clinically relevant scenario: a computational medical system must distinguish tumor-positive patches from normal tissues while identifying cases that are too uncertain to be safely handled without human expert review.

First, we apply different EDL methods for tumor detection. Besides classification accuracy (Acc, %), we report Diagnostic AUROC (DiagAUROC, %), which measures how well the predicted tumor probability ranks tumor-positive patches above normal patches. Error-Detection AUROC (ErrAUROC, %) evaluates whether the output uncertainty of the model ranks incorrectly classified cases above correctly classified ones.

However, in real workflows, it is rarely feasible for experts to manually re-check every patch processed by a machine learning model. We therefore select only the most uncertain 5% or 10% of test cases for human expert inspection. *Errors Captured*@k% (%) measures the fraction of all model errors that fall into this selected subset. *Missed Tumors Captured*@k% (%) is a safety-critical metric that measures the fraction of false-negative tumor cases captured by this subset. Finally, *Post-Referral Acc*@k% (%) estimates the accuracy after human review, assuming referred cases are corrected by experts. The value of k represents the selection rate.

Table 2 summarizes the performance of all compared methods, and CSEDL achieves the strongest performance on all the metrics. Furthermore, with only 10% of test cases selected and referred to human experts, CSEDL captures the most missed tumors in the selected subset and gets the best post-referral accuracy. This verifies the effectiveness of CSEDL and suggests that CSEDL can improve the performance of clinical tumor screening while only a small amount of human expert review is needed to recheck the model.

Table 3: Risk-aware trajectory forecasting under limited intervention budgets. Results are reported as mean ± standard deviation. The best results are in **bold**.

Method	Risk AUROC ↑	Risk AUPR ↑	BadFore.@5% ↑	BadFore.@10% ↑	Precision@5% ↑	Precision@10% ↑
EDL	63.91±0.86	18.62±1.14	10.41±1.02	17.83±1.26	27.67±1.74	23.70±1.31
$\mathcal{I}$-EDL	68.15±0.43	22.21±0.87	15.66±1.12	23.17±0.56	33.33±1.35	24.67±0.74
DAEDL	69.36±0.61	23.09±0.95	15.66±1.08	25.68±0.69	33.33±1.27	27.33±0.82
HEDL	66.23±0.77	19.97±0.58	13.99±1.03	19.42±0.92	29.78±1.15	20.67±0.49
$\mathcal{F}$-EDL	68.39±0.66	22.85±0.91	16.08±1.21	24.43±0.53	34.22±1.44	26.00±0.70
R-EDL	67.04±0.82	20.41±1.09	13.78±1.08	20.04±1.20	29.33±1.82	21.33±1.27
Re-EDL	67.38±0.79	20.99±1.05	14.20±0.97	20.87±1.18	30.22±1.63	22.37±1.23
CSEDL	**70.25±0.68**	**25.80±0.91**	**17.12±0.84**	**27.56±1.04**	**36.44±1.36**	**29.33±1.11**

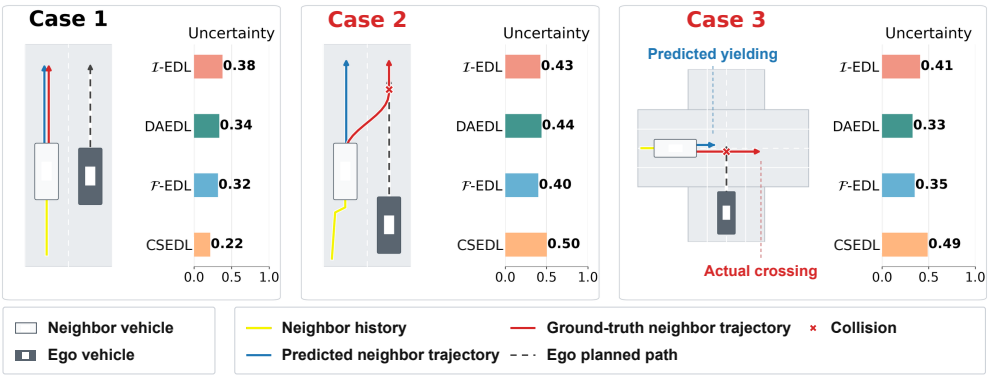


Figure 2: Forecast uncertainty in three driving scenarios evaluated by four EDL methods, in which no collision occurs in Case 1, while collisions occur in Case 2 and Case 3. **Case 1**: predicted and realized parallel motion agree. **Case 2**: predicted lane keeping becomes an actual lane change into the ego path. **Case 3**: predicted yielding becomes actual passage through the intersection.

4.4 Autonomous Trajectory Risk Forecasting

We evaluate CSEDL as a binary risk classifier for trajectory forecasts on the Argoverse 1 Motion Forecasting dataset, which includes observed vehicle motion histories and corresponding ground-truth future trajectories from real-world driving scenarios. In this setting, a trajectory predictor provides trajectory forecasts from motion histories. For a forecast generated by a shared trajectory predictor, we assign a high-risk label when the distance between its final position and that of the ground-truth trajectory exceeds 5 meters. Each EDL method is trained to identify high-risk forecasts from the shared trajectory predictor using observed vehicle motion histories, then we report the AUROC and AUPR of detecting high-risk forecasts (Risk AUROC and Risk AUPR).

In practice, when a forecast's risk level is uncertain, this forecast can be referred to a more reliable trajectory predictor, which is computationally intensive. So we rank all forecasts on the test set in descending order of the uncertainty produced by the EDL method and only select the top 5% or 10% for reassessment. BadFore.@$k\%$ measures the fraction of all high-risk forecasts captured in the selected ones and Precision@$k\%$ measures the fraction of selected forecasts that are high-risk, in which the value of k represents the selection rate.

Table 3 shows that CSEDL outperforms all the other baselines on both Risk AUROC and Risk AUPR. At the same selection budget, CSEDL also captures more high-risk forecasts with higher precision compared to these baselines. Figure 2 illustrates three synthetic driving scenarios. In Case 1, the neighboring vehicle continues straight alongside the ego vehicle, and its predicted and realized trajectories agree. In Case 2, the predictor forecasts lane keeping, but the neighbor changes lanes into the ego vehicle's planned path. In Case 3, the predictor forecasts yielding before the crossing, but the neighbor proceeds through the intersection. Compared with $\mathcal{I}$-EDL, DAEDL, and $\mathcal{F}$-EDL (the three strongest baselines in this task), CSEDL assigns the lowest uncertainty to the non-conflicting scenario (Case 1) and the highest uncertainty to the two potential-conflict scenarios (Cases 2 and 3) respectively, indicating CSEDL's ability to distinguish reliable forecasts from risky forecasts.

Table 4: Calibration performance of fine-tuned Qwen2.5-3B. Results are reported as mean ± standard deviation. The best results are in **bold** and the second best results are in underline.

Method	ARC-C			ARC-E		
	Acc ↑	ECE ↓	NLL ↓	Acc ↑	ECE ↓	NLL ↓
EDL	**82.46±0.20**	10.15±0.30	0.59±0.00	93.38±0.15	10.38±0.88	0.31±0.01
R-EDL	81.72±0.52	10.58±1.68	0.60±0.01	93.14±0.13	14.35±0.88	0.35±0.01
$\mathcal{I}$-EDL	81.77±0.64	12.33±2.51	0.62±0.01	93.28±0.26	13.59±1.02	0.33±0.01
IB-EDL	82.09±0.13	8.04±1.64	0.58±0.01	93.22±0.28	3.65±0.44	**0.26±0.00**
CSEDL	81.69±0.22	**6.30±1.49**	**0.56±0.01**	**93.46±0.09**	**3.13±0.56**	0.27±0.02
Method	OBQA			SciQ		
	Acc ↑	ECE ↓	NLL ↓	Acc ↑	ECE ↓	NLL ↓
EDL	88.67±0.81	11.11±0.20	0.44±0.02	93.83±0.15	10.26±0.10	0.29±0.00
R-EDL	88.40±0.20	12.14±0.65	0.46±0.01	93.63±0.47	8.15±4.50	0.26±0.05
$\mathcal{I}$-EDL	87.60±1.74	13.51±1.24	0.47±0.01	94.03±0.15	13.36±0.84	0.30±0.01
IB-EDL	88.33±0.23	**3.15±0.98**	0.41±0.02	**94.20±0.26**	3.83±0.44	0.21±0.01
CSEDL	**88.80±0.87**	4.86±1.09	**0.38±0.02**	94.07±0.23	**2.54±0.40**	**0.20±0.00**

4.5 Calibrating Confidence for Large Language Models

We further apply CSEDL to multiple-choice question answering of LLMs to calibrate its confidence of outputs. We perform supervised fine-tuning on the ARC-C, ARC-E, OBQA and SciQ datasets. All methods share the same Qwen2.5-3B backbone. For each multiple-choice question, we select the answer $\hat{y} = \arg\max_k p_k$ and define its confidence as $c = \max_k p_k$, where $k \in [1, 2, ..., K]$. Then we compare CSEDL against EDL, R-EDL and $\mathcal{I}$-EDL using classification accuracy (Acc), expected calibration error (ECE), and negative log-likelihood (NLL) to evaluate their calibration performance.

As shown in Table 4, CSEDL achieves the best performance on 66.67% (8 out of 12) cases and second best on 25.00% (3 out of 12) cases. These results show that CSEDL can improve LLM calibration performance while keeping competitive classification accuracy and negative log-likelihood. Moreover, although IB-EDL is specially designed for LLMs calibration, CSEDL outperforms it on 66.67% cases. This proves the generalization ability of our proposed method, i.e., performing well on different tasks without additional adaption.

4.6 Ablation Study

We examine the individual and joint contributions of uncertainty supervision $\mathcal{L}_{\text{unc}}$ and KL regularization $\mathcal{L}_{\text{kl}}$ on CIFAR-10, using SVHN, CIFAR-100, and GTSRB as OOD datasets and OOD AUROC of them is reported. The four variants share the same architecture and differ only in whether the corresponding loss coefficients are set to zero. As shown in Table 5, the complete model achieves the highest OOD AUROC across the three datasets and removing any term weakens the performance in both ID Acc and OOD AUROC, especially the latter. These results suggest complementary roles: uncertainty supervision supplies a sample-wise ranking signal, while KL regularization constrains non-target evidence and further improves detection when combined with this supervision.

Table 5: Ablation study on CIFAR-10 dataset. The best results are in **bold**.

$\mathcal{L}_{\text{unc}}$	$\mathcal{L}_{\text{kl}}$	ID Acc	SVHN	CIFAR-100	GTSRB
×	×	87.31	50.00	49.98	50.01
×	✓	89.58	63.47	60.39	55.81
✓	×	89.54	76.74	71.25	80.22
✓	✓	**92.50**	**97.01**	**87.59**	**90.68**

5 Conclusion

We proposed CSEDL, a dual-head evidential learning approach that combines class-direction supervision with uncertainty supervision. The direction head learns class predictions, while the uncertainty head is trained using ranks of nonconformity scores. Their outputs are then reconstructed as the Dirichlet parameters to form final predictions. Experiments on multiple tasks also demonstrate the effectiveness and practical value of our proposed method.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with grammar, spelling, and sentence clarity. We have reviewed all AI-assisted work and bear full responsibility for all content generated by the LLM. The generative AI contributed in no other way to the paper. The research idea, experiment design, and all other content were developed and completed by ourselves.

ETHICS STATEMENT

We confirm that our work adheres to the ICLR Code of Ethics. The research utilizes publicly available datasets containing no personally identifiable information. As a strictly technical contribution, we foresee no ethical issues or potential for misuse.

REPRODUCIBILITY STATEMENT

The complete code for our proposed method is provided in the Supplementary Material to facilitate the reproduction of our results. Our experimental implementation is also described in Section 4.1.

REFERENCES

Rina Foygel Barber and Ryan J. Tibshirani. Unifying different theories of conformal prediction. *Electronic Journal of Statistics*, 20(1):1428–1474, 2026.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 1613–1622, 2015.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Computer Vision – ECCV 2014*, volume 8694, pp. 446–461, 2014.

Ming-Fang Chang, John W. Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3D tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in neural information processing systems*, 33:1356–1367, 2020.

Bertrand Charpentier, Oliver Borchert, Daniel Zügner, Simon Geisler, and Stephan Günnemann. Natural posterior network: Deep bayesian uncertainty for exponential family distributions. *arXiv preprint arXiv:2105.04471*, 2021.

Mengyuan Chen, Junyu Gao, and Changsheng Xu. R-edl: Relaxing nonessential settings of evidential deep learning. In *International Conference on Learning Representations*, volume 2024, pp. 35563–35588, 2024.

Mengyuan Chen, Junyu Gao, and Changsheng Xu. Revisiting essential and nonessential settings of evidential deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10): 8658–8673, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Danruo Deng, Guangyong Chen, Yang Yu, Furui Liu, and Pheng-Ann Heng. Uncertainty estimation by fisher information-based evidential deep learning. In *International conference on machine learning*, pp. 7596–7616, 2023.

Robert Duin. Multiple features. UCI Machine Learning Repository, 1998.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *International Conference on Machine Learning*, 2017.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: Communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4:4, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7:17816, 2017.

Yawei Li, David Rügamer, Bernd Bischl, and Mina Rezaei. Calibrating llms with information-theoretic evidential deep learning. In *International Conference on Learning Representations*, 2025.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Rowan McAllister, Yarin Gal, Alex Kendall, Mark van der Wilk, Amar Shah, Roberto Cipolla, and Adrian Weller. Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 4745–4753, 2017.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Jingen Qu, Yufei Chen, Xiaodong Yue, Wei Fu, and Qiguang Huang. Hyper-opinion evidential deep learning for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37: 84645–84668, 2024.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, 2018.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *Proceedings of the International Joint Conference on Neural Networks*, pp. 1453–1460, 2011.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. In *Medical Image Computing and Computer Assisted Intervention*, volume 11071, pp. 210–218, 2018.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

Taeseong Yoon and Heeyoung Kim. Uncertainty estimation by density aware evidential deep learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *International Conference on Machine Learning*, volume 235, pp. 57217–57243, 2024.

Taeseong Yoon and Heeyoung Kim. Uncertainty estimation by flexible evidential deep learning. *Advances in Neural Information Processing Systems*, 38:118601–118641, 2026.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

TABLE OF CONTENTS FOR THE APPENDIX

- A: Overview of CSEDL 14
- B: Related Works of Evidential Deep Learning 14
- C: Pseudo-code of CSEDL 14
- D: Detailed Introduction of datasets 15
- E: Trusted Multi-View Fusion under Missing Views 16
- F: Sensitivity Analysis on maximum cache size $M_{\max}$ 16
- G: Further Analysis on Evolution of Uncertainty Distribution 17

A Overview of CSEDL

CSEDL is designed as a dual-head model as illustrated in Figure 3. The direction head is supervised by the ground-truth class label, whereas the uncertainty head is supervised by the empirical rank of the nonconformity score. Thus, class labels guide class prediction, while nonconformity score provides an additional training signal for uncertainty. The two outputs are then combined to reconstruct the Dirichlet parameters to obtain final prediction.

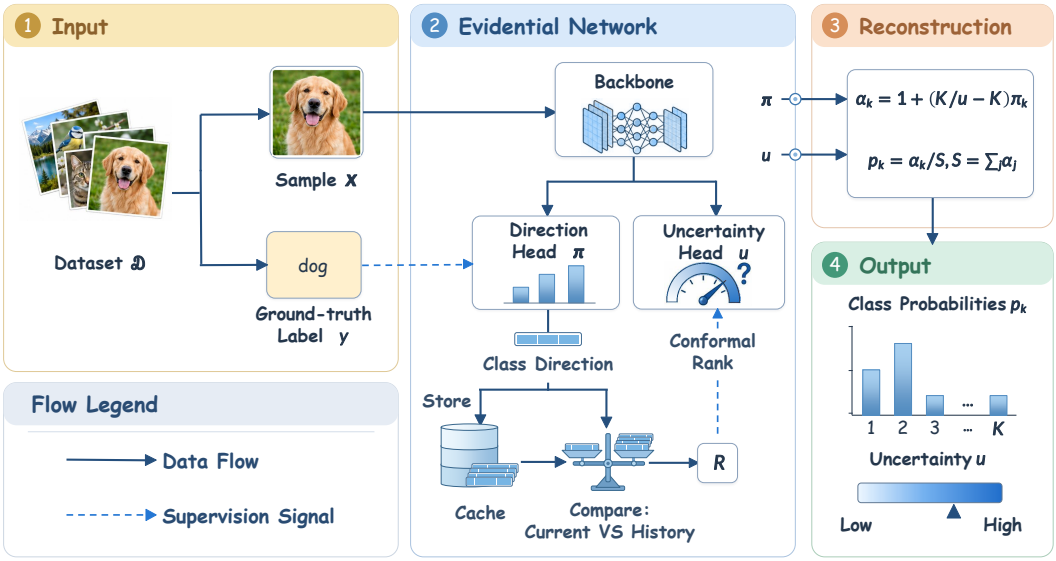


Figure 3: Overview of CSEDL with Dual-Head Learning and Separate Supervision.

B Related Works of Evidential Deep Learning

- Fisher Information-based Evidential Deep Learning ($\mathcal{I}$-EDL) (Deng et al., 2023) introduces the Fisher information matrix to characterize the informativeness of evidence and dynamically reweight the learning objective, together with a PAC-Bayesian regularization scheme.
- Density Aware Evidential Deep Learning (DAEDL) (Yoon & Kim, 2024) incorporates feature-space density into evidential prediction and introduces an alternative parameterization of the Dirichlet concentration to improve both OOD detection and classification.
- Hyper-opinion Evidential Deep Learning (HEDL) (Qu et al., 2024) extends conventional multinomial opinions to hyper-opinions, jointly modeling sharp evidence for individual classes and vague evidence shared by multiple classes.
- Flexible Evidential Deep Learning ($\mathcal{F}$-EDL) (Yoon & Kim, 2026) replaces the conventional Dirichlet assumption with a more expressive flexible Dirichlet distribution, aiming to improve uncertainty generalization under complex distributional conditions.
- Relaxed Evidential Deep Learning (R-EDL) (Chen et al., 2024) revisits the conventional EDL formulation by making the prior weight adjustable and removing the variance-minimizing term from the training objective.
- Re-EDL (Chen et al., 2025) further identifies the conventional KL-divergence regularizer as a nonessential component and simplifies evidential optimization by retaining the projected probability while removing both the variance-minimizing and KL-divergence terms.

C Pseudo-code of CSEDL

Algorithm 1 Pseudo-code of CSEDL

1: Initialize the CSEDL model and historical nonconformity score cache $\mathcal{M}$
2: **for** each training mini-batch $B = \{(x_i, y_i)\}_{i=1}^{b}$ **do**
3: Compute $\boldsymbol{\pi}_\theta(x_i)$, $q_\theta(x_i)$, $u_\theta(x_i)$, and the reconstructed Dirichlet parameters
4: Compute detached nonconformity scores $Z_i = 1 - \mathrm{sg}[p_{\theta,y_i}(x_i)]$
5: Estimate $\widehat{R}_i$ using the historical score cache
6: Compute $\mathcal{L}_{\mathrm{dir}}$, $\mathcal{L}_{\mathrm{unc}}$, and $\mathcal{L}_{\mathrm{kl}}$
7: Update model parameters using Eq. (16)
8: Append the detached current scores to $\mathcal{M}$
9: **end for**

D Detailed Introduction of Datasets

For OOD detection, we choose CIFAR-10 as the in-distribution (ID) dataset, SVHN, CIFAR-100, GTSRB, LFWPeople, Places365 and Food101 as the out-of-distribution (OOD) dataset.

- CIFAR-10 (Krizhevsky, 2009) contains 32×32 RGB images from ten common object categories, including animals and vehicles.
- SVHN (Netzer et al., 2011) contains house-number digits collected from Google Street View images. Its cropped-digit version provides 32×32 RGB images covering digits zero through nine.
- CIFAR-100 (Krizhevsky, 2009) contains 32×32 RGB images from 100 classes grouped into 20 superclasses. Its categories include animals, plants, household objects, and scenes.
- GTSRB (Stallkamp et al., 2011) contains cropped RGB images of German traffic signs from 43 classes, with variations in image resolution, viewpoint, and illumination.
- LFWPeople (Huang et al., 2007) provides identity-labeled face photographs from Labeled Faces in the Wild (LFW). The images depict faces under unconstrained conditions, with variations in pose, illumination, and facial expression.
- Places365 (Zhou et al., 2018) is a scene-classification dataset covering 365 categories of indoor and outdoor environments. Its images capture scene-level appearance and spatial context.
- Food101 (Bossard et al., 2014) contains food photographs from 101 categories, with substantial variation in food appearance, viewpoint, and illumination.

For tumor referral prediction, we use the PatchCamelyon (PCam) (Veeling et al., 2018) dataset. PCam is a binary histopathology benchmark consisting of RGB lymph-node tissue patches labeled according to the presence of metastatic tumor tissue.

For trajectory risk forecasting, we use the Argoverse 1 Motion Forecasting dataset (Chang et al., 2019). It contains real-world vehicle trajectories collected in urban driving scenarios.

For trusted multi-view classification, we use the Handwritten (Duin, 1998) multi-view benchmark. Each sample in this dataset is represented by six heterogeneous views, including Fourier coefficients, profile correlations, Karhunen–Loève coefficients, pixel averages, Zernike moments, and morphological features.

For large language model calibration, we choose ARC-C, ARC-E, OBQA and SciQ as datasets.

- ARC-C (Clark et al., 2018) contains challenging multiple-choice science questions from grade-school examinations. These questions were selected because both retrieval-based and word co-occurrence baselines failed to answer them correctly.
- ARC-E (Clark et al., 2018) contains the easier subset of the AI2 Reasoning Challenge, sharing the same grade-school science domain and multiple-choice format as ARC-C. Its questions were answered correctly by at least one of the two baseline methods used to partition the benchmark.
- OBQA (Mihaylov et al., 2018), short for OpenBookQA, contains elementary-level science questions with four answer options each. It is designed to assess reasoning that combines scientific facts with commonsense knowledge.

- SciQ (Welbl et al., 2017) contains crowdsourced multiple-choice science questions covering subjects such as physics, chemistry, and biology. Each question has four answer options, and most questions include a supporting passage providing evidence for the correct answer.

E TRUSTED MULTI-VIEW FUSION UNDER MISSING VIEWS

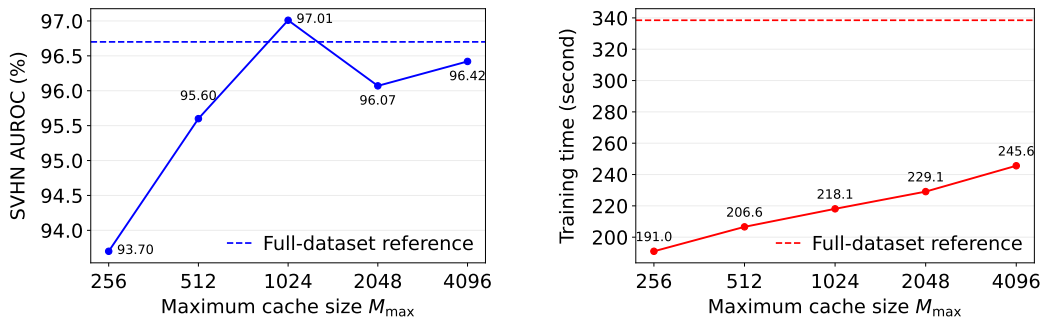
We evaluate CSEDL in a trusted multi-view classification setting using the Handwritten multi-view benchmark. Each sample is represented by six heterogeneous views. Trusted multi-view classification (TMC) aims to improve prediction reliability by accounting for the uncertainty of each view during fusion (Han et al., 2021). Following the TMC protocol, evidential heads on each view produce Dirichlet opinions and the view-level opinions are fused by Dempster–Shafer evidence fusion. We replace the conventional evidential head in each view branch with different evidential parameterizations while keeping the view encoders and the fusion rule unchanged. This tests whether a model can maintain reliable predictions when part of the input evidence is unavailable. For each model on every missing rate, we report classification accuracy (Acc, %), 15-bin expected calibration error (ECE, %) and negative log-likelihood (NLL).

Table 6: Missing-view robustness on the Handwritten multi-view dataset under different missing rates. Results are reported as mean ± standard deviation. The best results are in **bold**.

Method	Missing 10%			Missing 30%			Missing 50%		
	Acc ↑	ECE ↓	NLL ↓	Acc ↑	ECE ↓	NLL ↓	Acc ↑	ECE ↓	NLL ↓
EDL	97.55±0.48	3.19±0.37	0.1045±0.08	96.65±0.30	6.83±0.78	0.1841±0.09	94.35±0.34	13.44±0.08	0.3336±0.12
$\mathcal{I}$-EDL	97.55±0.33	13.37±0.63	0.2136±0.08	96.10±0.51	20.11±0.91	0.3285±0.10	94.65±0.60	28.46±1.34	0.5113±0.13
DAEDL	68.15±0.86	17.12±0.82	0.9731±0.11	56.00±2.02	14.76±0.69	1.2295±0.13	43.80±1.26	11.90±0.53	1.5112±0.15
HEDL	96.90±0.30	6.30±0.32	0.1175±0.06	96.10±0.44	12.48±0.58	0.2172±0.08	94.40±0.73	20.99±1.02	0.3852±0.11
$\mathcal{F}$-EDL	96.95±0.29	1.90±0.18	0.0825±0.05	96.85±0.37	4.07±0.24	0.1084±0.06	94.05±0.60	3.25±0.31	0.1802±0.08
R-EDL	97.25±0.42	9.22±0.82	0.1696±0.07	95.75±0.45	14.77±0.24	0.2844±0.09	91.60±0.46	20.78±0.59	0.4725±0.12
Re-EDL	97.15±0.46	5.52±0.14	0.1271±0.06	96.20±0.58	10.58±0.63	0.2218±0.09	92.95±0.56	16.38±0.74	0.3854±0.11
CSEDL	**97.75±0.27**	**1.81±0.19**	**0.0817±0.05**	**96.90±0.54**	**2.37±0.10**	**0.1076±0.06**	**94.75±0.45**	**2.77±0.28**	**0.1707±0.08**

Table 6 shows that CSEDL consistently improves trusted multi-view fusion under all missing-view rates in all three metrics compared to other baselines. It is also notable that even when 50% of the views are missing, CSEDL still reduces ECE to 2.77% and NLL to 0.1707. These results indicate that CSEDL controls the evidence strength separately from the direction prediction effectively, so the fused opinion remains substantially better calibrated even when half of the views are unavailable.

F SENSITIVITY ANALYSIS ON MAXIMUM CACHE SIZE $M_{\max}$



(a) SVHN AUROC under different cache sizes

(b) Training time under different cache sizes

Figure 4: Parameter sensitivity of the cache size $|\mathcal{M}|$. The dashed horizontal line in each panel denotes the Full-dataset reference strategy, which recomputes the nonconformity scores over the training set at the beginning of each epoch to construct $\widehat{F}_t$. Left: SVHN OOD-detection AUROC with CIFAR-10 as the in-distribution dataset. Right: corresponding training time.

We further investigate the sensitivity of CSEDL to the maximum cache size $M_{\max}$ and compare the proposed nonconformity score cache with a more expensive full-dataset reference strategy mentioned in Section 3.3. Specifically, we recompute the nonconformity scores over the entire CIFAR-10 training set at the beginning of each epoch and construct the corresponding empirical score distribution $\widehat{F}_t$ for this strategy.

Figure 4 compares OOD detection performance and training time on CIFAR-10 versus SVHN. The full-dataset reference requires substantially more training time without improving AUROC over the online cache with $M_{\max} = 1024$. For the online cache, AUROC increases from 93.70% to 97.01% as $M_{\max}$ grows from 256 to 1024, but further increasing the cache size does not improve performance beyond this peak. Meanwhile, training time increases steadily with cache size, rising from 191.0 seconds at $M_{\max} = 256$ to 245.6 seconds at $M_{\max} = 4096$. Balancing detection performance and training efficiency, we finally choose $M_{\max} = 1024$ because it achieves the highest AUROC among the tested cache sizes while avoiding additional runtime.

G Further Analysis on Evolution of Uncertainty Distribution

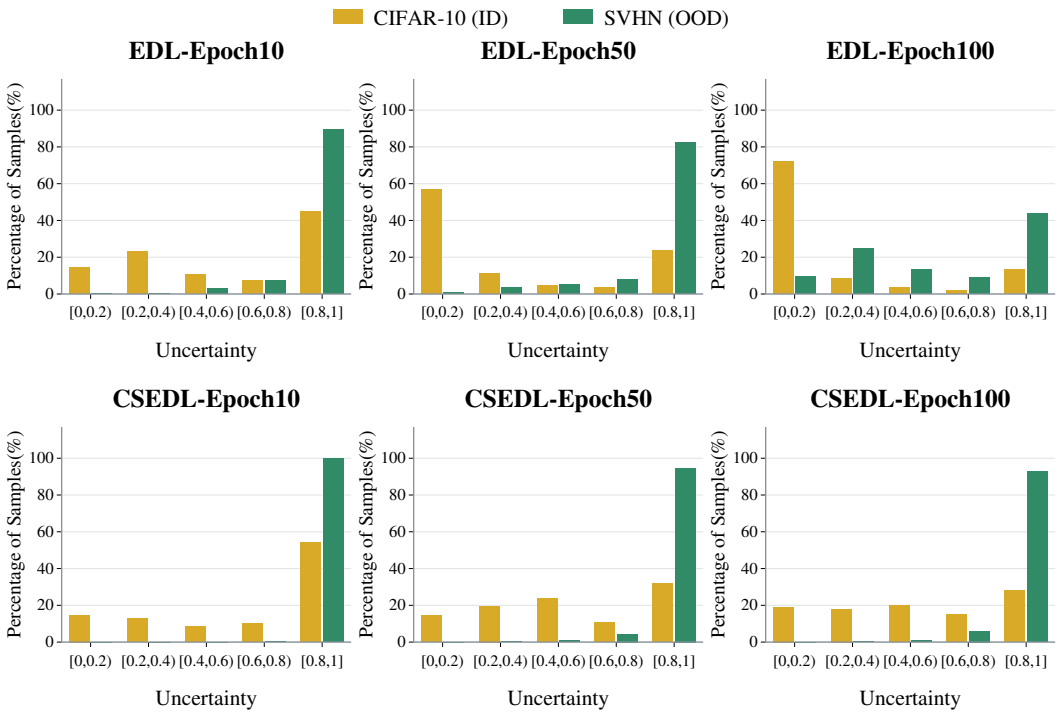


Figure 5: Comparison of EDL and CSEDL in distributions of uncertainty on CIFAR-10 (ID, yellow) and SVHN (OOD, green) at three training epochs (10, 50 and 100). Bars report the percentage of samples in each uncertainty range.

We delve into the training process of conventional EDL and CSEDL and inspect the trend of uncertainty distribution of ID and OOD samples. Figure 5 shows the uncertainty distributions in CIFAR-10 (in-distribution (ID), yellow) and SVHN (out-of-distribution (OOD), green) at three training epochs (10, 50, and 100). The uncertainty is divided into five bins, and percentage of samples in each uncertainty bin is plotted. It can be observed that in conventional EDL, although uncertainty decreases on in-distribution (ID) samples as training progresses, a similar shift also occurs on OOD inputs. This proves that evidence can accumulate excessively on OOD inputs during training, consequently leading to inaccurate uncertainty. In CSEDL, uncertainty of ID samples also decreases, while the majority of OOD samples keep high uncertainty in $[0.8, 1]$, showing the effectiveness of our method in quantifying uncertainty.