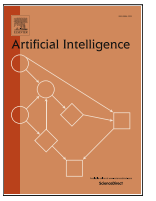
Artificial Intelligence 360 (2026) 104608



Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint



Evidence-guided learning against noisy and sparse labels on graphs

Siyu Yi[a], Wei Zhang[b], Zhengyang Mao[c], Yongdao Zhou[d], Ziyue Qiao[e], Li Shen[f], Dacheng Tao[g], Jiancheng Lv[b], Wei Ju[h,*]

[a] School of Mathematics, Sichuan University, Chengdu, 610065, China
[b] College of Computer Science, Sichuan University, Chengdu, 610065, China
[c] School of Computer Science, Peking University, Beijing, 100871, China
[d] School of Statistics and Data Science, Nankai University, Tianjin, 300071, China
[e] School of Computing and Information Technology, Great Bay University, Dongguan, 523000, China
[f] School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Guangzhou, 518107, China
[g] School of Computer Science, The University of Sydney, New South Wales, 2006, Australia
[h] School of Artificial Intelligence, Sichuan University, Chengdu, 610065, China

ARTICLE INFO

Keywords:
Graph neural networks
Label noise
Uncertainty estimation
Multi-expert learning

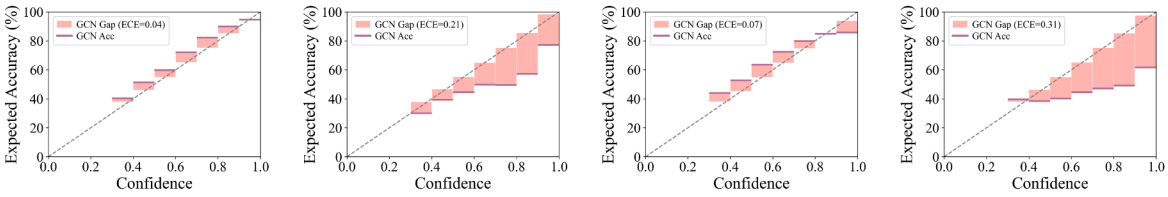
ABSTRACT

With the tremendous expansion of graph-structured data, graph neural networks (GNNs) have emerged as the essential tools to tackle graph-related problems. Existing methods generally assume an ideal setting where all class labels are clean. However, due to the label noise and expensive annotation cost, many real-world graph labels are incorrect and scarce. Directly training GNNs under such supervision may lead to overfitting and poor generalization. Despite the effectiveness achieved by existing approaches for handling noisy labels in vision domains, the study on graphs remains largely under-explored. In this paper, we explore a significant yet challenging scenario where nodes on the graph are sparsely and noisily labeled. To address these issues, we propose an evidence-guided framework (EvGNN) to mitigate the negative impact of noisy labels and enhance the utilization of unlabeled nodes on graphs in a multi-expert fashion. Specifically, EvGNN leverages the idea of subjective logic guided by the Dempster-Shafer evidence theory to provide evidence and quantify uncertainty for each expert. Then a novel fusion strategy is developed to overcome highly conflicting phenomena among experts. Moreover, to boost supervisory signals on the graph, we further propose consistency learning and pseudo-labeling as regularizations to fully leverage the unlabeled nodes, thus providing auxiliary supervision against sparse labels. We show the effectiveness of our proposed method through extensive experiments on real-world datasets.

Graph-structured data can be found in many real-world situations and typically serve as a common language for representing the interactions between various entities, such as social networks, citation networks, molecular graphs and knowledge graphs. To analyze such relational data, which are pervasive in artificial intelligence applications, there has been a significant amount of interest in developing models that can be applied to a wide range of domains and tasks. Among its solutions, graph neural networks (GNNs) [1–3] have achieved remarkable success and emerged as powerful approaches recently. The key idea of GNNs is to learn an expressive node representation by propagating and transforming representations of its neighbors following a message-passing mechanism [4].

* Corresponding author.
E-mail address: juwei@scu.edu.cn (W. Ju).

https://doi.org/10.1016/j.artint.2026.104608
Received 17 April 2025; Received in revised form 25 August 2026; Accepted 26 August 2026
Available online 2 September 2026
0004-3702/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.



(a) 20% Noise, 10% Labels (b) 20% Noise, 1% Labels (c) 40% Noise, 10% Labels (d) 40% Noise, 1% Labels

Fig. 1. Motivating experiment on Pubmed under uniform label noise and sparse supervision. ECE measures the discrepancy between prediction confidence and empirical accuracy, where a lower value indicates better calibration.

Despite the encouraging performance of existing GNNs in capturing attributive and structural information on graphs, most existing approaches typically assume that the training labels are clean and sufficient. In fact, for many real-world graphs and applications, the labels are often noisy and extremely limited. For example, on the one hand, when annotating node labels in crowdsourcing, since the majority of people are not domain experts, their confidence in label annotation is relatively low. Such subjective judgments inevitably lead to labeling errors, resulting in the introduction of noisy labels. Moreover, previous studies have shown that noisy labels can severely weaken the robustness of deep learning models and lead to limited generalization [5]. On the other hand, in many high-precision fields, such as biomedicine [6], labeling a molecule or protein is very time-consuming and expensive, and it also requires the involvement of domain experts. This dilemma results in label sparsity. As such, how to effectively resist noisy labels and overcome the sparsity of the labels becomes a critical research problem.

Actually, there are a large number of existing ideas in vision domains to address the aforementioned issues, which can be categorized into three groups: *sample selection*, *loss correction* and *label correction*. The sample selection strategy [7–9] aims to select clean labels during the training process to optimize deep learning models. The loss correction strategy [10–12] aims to correct the impact of label noise on risk minimization by adjusting the loss of training samples. Label correction [13–15] is a very intuitive solution that can identify data with potentially incorrect labels and correct them to achieve reliable training. To adapt these ideas well to graph-structured data, indeed, there are only a handful of works on graphs [16–19]. NRGNN [16] establishes connections between unlabeled and labeled nodes that share similar features to enhance the predictive accuracy and reliability of label information. RTGNN [17] adopts two peer GNNs to divide labels based on the loss to identify labels that are likely to be noisy and further rectify them. CR-GNN [19] enhances robust graph representation learning by unifying contrastive neighbor alignment, dynamic reliable-node selection, and cross-space consistency optimization.

Thus, learning against label noise on graphs remains largely unexplored. Moreover, most existing methods still suffer from two key limitations: **(i) Inability to provide evidence and quantify uncertainty of the prediction for each node.** While GNNs have shown remarkable success of predictive accuracy, their reliability cannot always be guaranteed. Research has demonstrated that traditional neural networks tend to be overconfident [20–22], which poses a challenge to some failure-sensitive applications that require high reliability, such as disease diagnosis [23] and drug recommendation [24]. Moreover, noisy labels can also affect the model robustness during the training process and inevitably result in incorrect predictions with high confidence [7,25]. Therefore, it is crucial for GNNs to evaluate their predictive uncertainty and avoid incorrect judgments. **(ii) Suffering from labeled data scarcity and insufficient usage of unlabeled nodes**. Since the annotation process is labor-intensive and expensive, many real-world graphs are often sparsely labeled, typically leading to unsatisfactory performance. Besides, though GNNs implicitly utilize the features of unlabeled nodes via message-passing mechanisms [4], their potential value has not been fully explored, whose inherent attributive and structural information can serve as a regularizer to guide the GNN learning process. To empirically illustrate the influence of both label noise and label sparsity on prediction confidence, we conduct a motivating experiment using a vanilla GCN with cross-entropy loss on Pubmed under uniform label noise. As shown in Fig. 1, when the noise rate increases or label rate decreases, the calibration gap becomes more pronounced and the empirical accuracy tends to fall further below the corresponding prediction confidence. Consequently, the expected calibration error (ECE) also increases, revealing substantial overconfidence in the predictions of a conventional GCN, especially when noisy labels and sparse supervision coexist. Therefore, it is essential to quantify the uncertainty of the predictions and fully leverage unlabeled nodes to enhance the robustness of GNNs against label noise and sparsity.

Toward this end, we propose an evidence-guided framework named EvGNN for robust training of GNNs. The key idea of EvGNN is to effectively reduce the negative effects of noisy labels and sufficiently improve the utilization of unlabeled nodes. To achieve this goal, we build our approach upon an evidence-guided multi-expert framework. Uncertainty quantification is employed to characterize prediction reliability by explicitly modeling evidence/belief and uncertainty in the output space, while diverse experts capture complementary predictive patterns from different perspectives in the model space. By integrating the resulting evidential opinions, the framework reduces the risk of being jointly misled by noisy labels and yields more reliable predictions than a single model. Specifically, EvGNN first introduces subjective logic supported by the Dempster-Shafer evidence theory [26] to produce class-wise evidence and quantify uncertainty for each expert. The evidence parameterizes a Dirichlet distribution over class probabilities, allowing the model to represent both prediction confidence and uncertainty under noisy supervision. Based on the evidential opinions from multiple experts, we further design a conflict-aware decision fusion strategy, where expert conflict and information volume are considered before applying Dempster's combination rule. It helps reduce the influence of unreliable or highly conflicting expert opinions.

To better exploit unlabeled nodes, EvGNN also performs representation-level fusion and aligns it with decision-level fusion through consistency learning, so that the two fusion views can regularize each other under label scarcity. Finally, high-confidence predictions from the fused evidence are selected as pseudo labels to provide additional supervision. Rather than treating these components as independent add-ons, EvGNN connects them through evidence and uncertainty modeling, forming a unified uncertainty-aware training framework for robust graph learning under noisy and sparse labels.

To summarize, the main contributions of our work are as follows:

- We explore a largely under-explored problem that trains GNNs against noisy and sparse labels, and innovatively incorporate the uncertainty and confidence into this setting to enhance the model robustness and generalization.
- We propose an evidence-guided framework to explicitly provide uncertainty of the nodes in a multi-expert framework against label noise and fully leverage the unlabeled nodes through consistency learning and pseudo-labeling as regularization against label sparsity.
- We conduct comprehensive experiments to evaluate the superiority of our proposed approach over state-of-the-art methods on various commonly used datasets.

1. Related work

1.1. Graph neural networks

Graph Neural Networks (GNNs) have been widely studied because of their ability to encode graph-structured data and have achieved promising performance in node classification [1,27], graph clustering [28–30], and graph classification [31,32]. Early spectral methods define graph convolution through graph Fourier bases [33], whereas later spatial methods update node representations by aggregating and transforming neighborhood information under the message-passing paradigm [4]. Representative spatial GNNs differ mainly in how they construct and aggregate neighborhood information. GCN [1] applies normalized graph convolution, whereas GraphSAGE [34] learns inductive aggregators that can generalize to unseen nodes. GAT [35] subsequently introduces neighbor-specific attention weights. From a theoretical perspective, GIN [36] examines how the choice of aggregation function determines the expressive power of a GNN. More recently, graph Transformer models have introduced global or structure-aware attention to graphs, for example by learning meta-path-like relations [37] or encoding centrality, spatial, and edge information into Transformer attention [38]. These developments show that GNN backbones have evolved well beyond plain graph convolution and can capture different local and global dependencies.

Most of these architectures, however, are designed under the assumption that the available labels are reliable enough to supervise representation learning [39]. Under noisy and sparse supervision, stronger propagation or attention alone cannot ensure robustness. The classifier may still memorize corrupted labels, while message passing can spread their influence to neighboring unlabeled nodes. Hence, to improve robustness under noisy and sparse supervision, our proposed EvGNN incorporates uncertainty estimation on the class probabilities from a Bayesian perspective against label noise, and sufficiently explores the unlabeled nodes against label scarcity.

1.2. Neural networks with noisy labels

It has been demonstrated that standard deep neural networks may overfit noisy labels and consequently suffer from poor generalization [5]. Extensive studies in conventional data domains have developed robust training strategies that reduce the influence of unreliable supervision and prevent models from fitting corrupted annotations [8,9,11,12,14,15]. For example, Decoupling [9] involves training two networks to select clean samples and updates each other with the clean samples they generated, and Hu et al. [14] uses a self-training approach that involves alternating between correcting the labels and training a deep neural network on the noisy-labeled data. Their effectiveness in standard classification settings shows that controlling how noisy annotations contribute to model training can substantially reduce overfitting and improve generalization. However, these strategies are mostly developed for independent and identically distributed samples. When they are transferred to graphs, they may ignore the dependency between node features, labels, and graph structure, which is crucial because message passing can amplify both useful supervision and erroneous labels.

Graph-specific noisy-label learning has therefore attracted increasing attention [40–45]. PI-GNN [40] formulates robust node classification from a pairwise perspective. It estimates whether two nodes belong to the same class and uses the resulting pairwise labels to regularize the node classifier. GNN Cleaner [41] generates neighborhood-based pseudo-labels through modified label propagation and adaptively corrects unreliable annotations during training. More recent methods make direct use of graph topology when identifying or correcting noisy labels. TSS [42] incorporates topological information into sample selection so that informative nodes near class boundaries are not discarded merely because they are difficult to classify. R^2LP [43] accommodates different levels of graph heterophily by iteratively reconstructing the graph, rectifying labels through label propagation, and retaining high-confidence labels. ERASE [44] first performs structural denoising and then combines the corrected labels with prototype-based pseudo-labels to learn error-tolerant representations through coding rate reduction. NoisyGL [45] establishes a unified benchmark with consistent data splits, preprocessing procedures, and evaluation protocols, allowing graph noisy-label methods to be compared under controlled settings.

Existing graph-specific approaches address label noise at different stages of graph learning, either correcting unreliable labels before they affect optimization or reducing noise propagation through sample selection, graph reconstruction, and robust representation learning. However, correction and pseudo-labeling depend on the accuracy of intermediate predictions, while other approaches

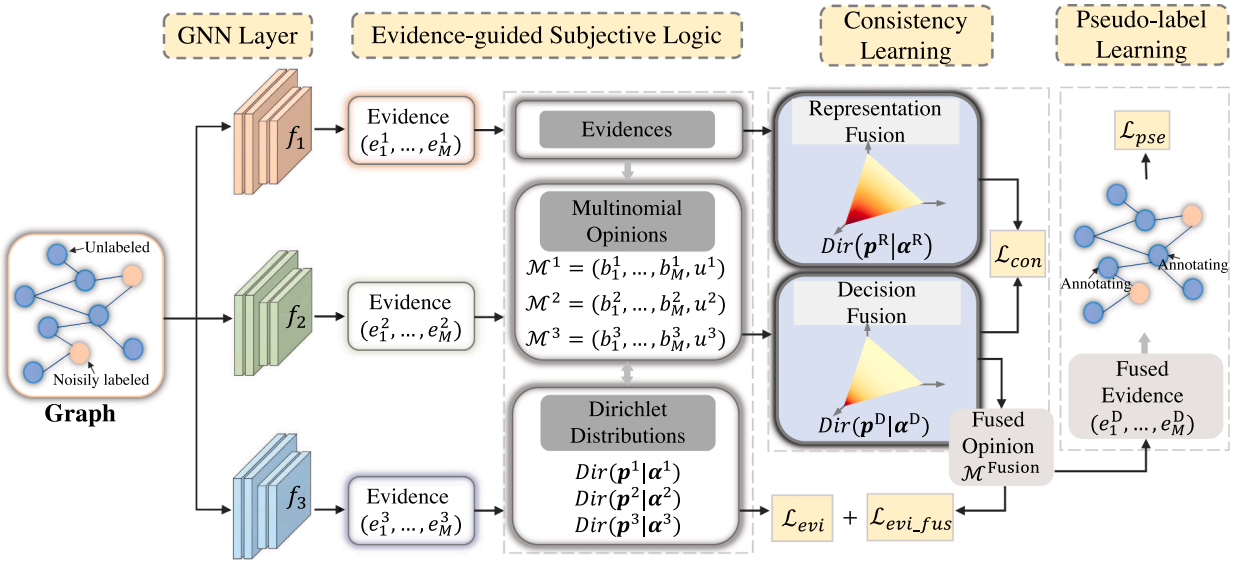


Fig. 2. Framework overview of the proposed method EvGNN.

generally treat predictions as deterministic and neither quantify uncertainty nor assess evidence strength and conflicts among multiple opinions. These limitations become more pronounced when labels are simultaneously noisy and sparse. Motivated by this gap, our EvGNN models class-wise evidence and uncertainty for each expert and performs conflict-aware fusion according to both disagreement and information volume. The fused evidence further guides consistency learning and pseudo-labeling, jointly reducing the influence of corrupted labels and improving the use of unlabeled nodes.

1.3. Uncertainty-based learning

Uncertainty estimation is crucial in revealing the reliability for neural networks, which often struggle to provide accurate predictions [46–50]. Bayesian neural networks and Gaussian process are two prominent techniques used to estimate uncertainty in deep learning, where Bayesian neural networks [51–53] leverage prior distributions on model parameters and Gaussian processes [54] defines a prior over functions to compute posterior uncertainty exactly [55]. However, these methods can be computationally costly and slow during inference. Orthogonal to these methods, subjective logic is another line that allows us to reason about uncertainty. It extends the Bayesian theory to handle subjective probabilities [56], which is guided by the Dempster-Shafer evidence theory (DST) [57]. DST excels at combining the evidences from multiple sources, which makes subjective logic more flexible compared to the Bayesian theory. Leveraging the theoretical foundation and adaptability of subjective logic, our work can directly provide estimates of uncertainty and predictive confidence.

2. Preliminaries

Notations. Given a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{X}, \mathbf{Y}_{\mathrm{L}}\}$ with N nodes and M classes. $\mathcal{V} = \mathcal{V}_{\mathrm{L}} \cup \mathcal{V}_{\mathrm{U}}$ is the node set of the graph, where $\mathcal{V}_{\mathrm{L}}(|\mathcal{V}_{\mathrm{L}}| = L)$ and $\mathcal{V}_{\mathrm{U}}(|\mathcal{V}_{\mathrm{U}}| = N - L)$ are the labeled and unlabeled node sets, respectively. $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set and $\mathbf{X} \in \mathbb{R}^{N \times d}$ is the node attribute matrix, where d is the dimention of the attributes. $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix, in which $\mathbf{A}_{ij} = 1$ if there is an edge between nodes i and j; $\mathbf{A}_{ij} = 0$ otherwise. The one-hot labels for the nodes in $\mathcal{V}_{\mathrm{L}}$ are denoted by $\mathbf{Y}_{\mathrm{L}} = \{\mathbf{y}_1, \ldots, \mathbf{y}_L\} \in \{0, 1\}^M$, which are limited and partially polluted by noise.

Problem Definition. Given a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{X}, \mathbf{Y}_{\mathrm{L}}\}$, where the observed labels $\mathbf{Y}_{\mathrm{L}}$ may be noisy and the labeled node set $\mathcal{V}_{\mathrm{L}}$ is sparse, we focus on semi-supervised node classification. The goal is to learn a robust GNN $\mathcal{F} : \mathcal{G} \to \hat{\mathbf{Y}}_{\mathrm{U}}$ that can effectively predict the labels $\hat{\mathbf{Y}}_{\mathrm{U}}$ of unlabeled nodes in $\mathcal{V}_{\mathrm{U}}$.

Graph Convolutional Network (GCN). GCN [1] is a neural network tailored for graph-structured data. It extends the conventional convolutional neural networks to operate within the graph domain, which learns node representations by propagating and aggregating information from a node and its neighbors. Mathematically, the $(c+1)$th layer of GCN typically has the form $\mathbf{H}^{(c+1)} = \sigma(\tilde{\mathbf{A}}\mathbf{H}^{(c)}\mathbf{W}^{(c)} + \mathbf{b}^{(c)})$, where $\tilde{\mathbf{A}} = \mathbf{D}^{-1/2}\bar{\mathbf{A}}\mathbf{D}^{-1/2}$ is the normalized adjacency matrix, $\mathbf{D}$ is the diagonal degree matrix with $\mathbf{D}_{ii} = \sum_{j=1}^{N} \bar{\mathbf{A}}_{ij}$, $\bar{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\mathbf{I} \in \mathbb{R}^{N \times N}$ is the identity matrix, $\sigma(\cdot)$ is an activation function (e.g., Tanh or ReLU), $\mathbf{H}^{(c)}$ is the learned embedding, and $\mathbf{W}^{(c)}$ and $\mathbf{b}^{(c)}$ are the trainable weight matrix and bias vector at the cth layer, respectively.

3. The proposed method

In this section, we present our framework EvGNN to learn a robust GNN against noisy and sparse labels. The framework mainly consists of four parts: evidence-based uncertainty quantification, decision fusion of multiple experts, representation fusion of multiple experts, and consistency and pseudo-label learning. Each component is introduced with a specific motivation. First, evidence-based learning with subjective logic [58,59] is used to model prediction reliability by quantifying the uncertainty caused by noisy labels. Second, we leverage multi-expert learning [7] to reduce the risk that a single model memorizes corrupted labels, since different experts can provide complementary evidential opinions. Third, decision-level fusion aggregates these expert opinions while considering both consensus and uncertainty, thereby improving the robustness of the final decision. Fourth, representation-level fusion provides another evidence view from expert embeddings, which is further aligned with decision-level fusion through consistency learning to exploit unlabeled nodes. Finally, pseudo-labeling uses high-confidence fused evidence to provide auxiliary supervision under label scarcity. Fig. 2 exhibits the framework overview of EvGNN.

3.1. Evidence-based uncertainty quantification

Evidence-based subjective logic is a probabilistic reasoning framework that explicitly accounts for epistemic uncertainty and source reliability [50]. It provides a principled mechanism for quantifying both the belief masses assigned to candidate classes and the uncertainty mass (i.e., belief vacuity), thereby forming a multinomial opinion. By explicitly modeling both class-specific beliefs and uncertainty, the resulting opinion provides a more informative characterization of predictive confidence than conventional point estimates. Such a property is particularly desirable in noisy-label settings, where misleading annotations may induce unreliable predictions. Therefore, the primary role of uncertainty quantification is to characterize prediction reliability under noisy supervision.

However, uncertainty quantification primarily focuses on calibrating prediction confidence and does not explicitly prevent the model from memorizing and overfitting noisy labels during training [7,25]. To address this limitation, we introduce a multi-expert learning framework. Owing to different parameter initializations, stochastic optimization trajectories, and the non-convex nature of the training objective, different experts can converge to distinct local optima even when trained on the same graph. Consequently, they may learn complementary predictive patterns and exhibit different sensitivities to noisy annotations. Such diversity reduces the possibility that all experts are simultaneously misled by the same noisy labels, indicating that the multi-expert design is not used merely for model averaging, but to provide diverse evidential opinions that can be further fused into a more reliable uncertainty-aware decision.

Formally, in the M-class node classification task under a K-expert framework, given any node on the graph, evidence-based subjective logic assigns a *belief* mass b^i_m to each class $m = 1, \dots, M$, and provides an overall *uncertainty* mass u^i for the ith expert. The $M + 1$ mass values (multinomial opinion) are all non-negative and sum up to one, i.e., these masses satisfy $\sum_{m=1}^{M} b^i_m + u^i = 1$. The calculation of belief and uncertainty relies on the *evidence* for different classes derived from the data, which can be acquired by a GCN based on node attributes and topology information, followed by a multi-layer perceptron (MLP) and an activation operation that projects the dimension to the class number M with non-negative entries, i.e.,

$$\mathbf{E}^i = \text{Softplus}(\text{MLP}(\sigma(\tilde{\mathbf{A}}\sigma(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}^{(1),i} + \mathbf{b}^{(1)})\mathbf{W}^{(2),i}) + \mathbf{b}^{(2)})),$$

where $\mathbf{E}^i \in \mathbb{R}^{N\times M}$ denotes the evidence matrix generated by the ith expert, the sth row $\mathbf{e}^i_s \in \mathbb{R}^M$ of $\mathbf{E}^i$ represents the evidence vector of the sth node, and its mth element e^i_{sm} quantifies the support from the observed data for assigning the sth node to class m. Without causing confusion, we omit the superscript s for e^i_{sm} in the following discussion. Then the belief b^i_m and uncertainty u^i can be calculated by

$$b^i_m = \frac{e^i_m}{S^i}, \quad u^i = \frac{M}{S^i}, \text{ with } S^i = \sum_{m=1}^{M}(e^i_m + 1) \text{ for } m = 1, \dots, M. \tag{1}$$

It is obvious that the uncertainty mass is negatively correlated with the total evidence (or belief masses). With the inclusion of uncertainty, such a paradigm is beneficial to avoid the false positive predictions [50] and incorporates network training and uncertainty quantification into a unified framework without additional computations.

Following subjective logic, the multinomial opinion induces a Dirichlet distribution with concentration parameters $\{\alpha^i_m = e^i_m + 1\}_{m=1}^{M}$, whose density has the form

$$Dir(\mathbf{p}^i|\boldsymbol{\alpha}^\mathbf{i}) = \frac{\boldsymbol{\Gamma}(\mathbf{S^i})}{\prod_{\mathbf{m=1}}^{\mathbf{M}} \boldsymbol{\Gamma}(\boldsymbol{\alpha}^\mathbf{i}_\mathbf{m})} \prod_{\mathbf{m=1}}^{\mathbf{M}} (\mathbf{p}^\mathbf{i}_\mathbf{m})^{\boldsymbol{\alpha}^\mathbf{i}_\mathbf{m}-\mathbf{1}},$$

where $\mathbf{p}^i = (p^i_1, \dots, p^i_M)^\top$ is the class probability assignment satisfying $\sum_{m=1}^{M} p^i_m = 1$, $\boldsymbol{\alpha}^\mathbf{i} = (\boldsymbol{\alpha}^\mathbf{i}_\mathbf{1}, \dots, \boldsymbol{\alpha}^\mathbf{i}_\mathbf{M})^\top$, and $\Gamma(\cdot)$ is the gamma function. Such Dirichlet distribution characterizes the possible values of the class assignment $\mathbf{p}^i$ under the support of the belief for the ith expert, which is equivalent to the multinomial opinion [58,60], thereby measuring the uncertainty of the class assignment. Subsequently, we regard the Dirichlet distribution as the prior information for the GCN-based classifier, where the cross-entropy loss is commonly adopted. Then, following the Bayesian inference, the evidence-based supervised loss under the K-expert framework is formulated as

$$\mathcal{L}_{evi} = \sum_{i=1}^{K}\sum_{s=1}^{L}\int \left[\sum_{m=1}^{M} -y_{sm}\log(p^i_{sm})\right] Dir(\mathbf{p}^i_s|\boldsymbol{\alpha}^\mathbf{i}_\mathbf{s})\mathrm{d}\mathbf{p}^\mathbf{i}_\mathbf{s} = \sum_{\mathbf{i=1}}^{\mathbf{K}}\sum_{\mathbf{s=1}}^{\mathbf{L}}\sum_{\mathbf{m=1}}^{\mathbf{M}} \mathbf{y_{sm}}(\boldsymbol{\psi}(\mathbf{S}^\mathbf{i}_\mathbf{s}) - \boldsymbol{\psi}(\boldsymbol{\alpha}^\mathbf{i}_\mathbf{sm})), \tag{2}$$

where $\mathbf{y}_s = (y_{s1}, \ldots, y_{sM})^\top$ is the annotated one-hot label for the sth node in $\mathcal{V}_L$, $\psi(\cdot)$ is the digamma function, $S_s^i = \sum_{m=1}^M \alpha_{sm}^i$, $\mathbf{p}_s^i = (p_{s1}^i, \ldots, p_{sM}^i)^\top$ and $\boldsymbol{\alpha}_s^i = (\alpha_{s1}^i, \ldots, \alpha_{sM}^i)^\top$ are the class probability assignment and the concentration parameters of the Dirichlet distribution for the sth node from the ith expert, respectively. Compared with the classical cross-entropy, the Bayesian paradigm optimizes the network in the average sense of all possible class assignments, which can effectively alleviate the misguidance of noisy labels.

3.2. Decision fusion of multiple experts

The effectiveness of multi-expert learning relies on the diversity among experts. Although all experts are trained on the same graph, they may capture complementary information due to differences in parameter initialization and optimization trajectories [47, 61,62]. Specifically, for evidence and uncertainty learning, due to different optimization trajectories, different experts can converge to distinct evidence allocation patterns, leading to diverse multinomial opinions for the same node; for graph message passing, different parameter initializations may result in distinct learned weight matrices, thereby placing varying emphasis on local and long-range neighborhood information. Consequently, experts may exhibit different effective receptive fields and capture different structural characteristics of the graph. Such diversity enables experts to provide complementary predictive signals and uncertainty assessments. Therefore, to obtain more reliable and robust predictions, we introduce an elaborate decision fusion strategy that integrates beliefs and uncertainties across experts, allowing both consensus and uncertainty to contribute to the final decision.

Based on the obtained multinomial opinions $\{\mathcal{M}^i = \{b_1^i, \ldots, b_M^i, u^i\}\}_{i=1}^K$ from K experts for any node on the graph, a classic data fusion method is the Dempster's combination rule [63], which is defined as follows.

Definition 1. (Dempster's combination rule) Let $\mathcal{M}^i = \{b_1^i, \ldots, b_M^i, u^i\}$ and $\mathcal{M}^j = \{b_1^j, \ldots, b_M^j, u^j\}$ be two multinomial opinions from any two different experts. Dempster's rule of combination is defined as

$$\mathcal{M}^{\text{Binary}} = \{b_1, \ldots, b_M, u\} \triangleq \mathcal{M}^i \oplus \mathcal{M}^j, \tag{3}$$

where $b_m = (b_m^i b_m^j + b_m^i u^j + b_m^j u^i)/(1-C)$ for $m = 1, \ldots, M$, $u = u^i u^j/(1-C)$, $C = \sum_{p \neq q} b_p^i b_q^j$ is a measure of conflict degree between $\mathcal{M}^i$ and $\mathcal{M}^j$, and $1-C$ is normalization factor.

Further, given $K(>2)$ experts in a multi-expert framework, we apply the combination rule in a sequential manner to formulate the fused multinomial opinion as $\mathcal{M}^{\text{Multi}} \triangleq \mathcal{M}^1 \oplus \mathcal{M}^2 \cdots \oplus \mathcal{M}^K$.

However, the term $1-C$ in Eq. (3) overlooks conflicts entirely, assigning any mass function related to conflicts to an empty set. To illustrate, consider a simple case with two experts and three classes $\{C_1, C_2, C_3\}$. Suppose $\mathcal{M}^1 = \{0.99, 0, 0, 0.01\}$ and $\mathcal{M}^2 = \{0, 0.99, 0, 0.01\}$. Intuitively, the two experts are in strong conflict: one strongly supports C_1, the other strongly supports C_2. According to Dempster's rule, the conflict degree is $C = 0.99 \times 0.99 = 0.9801$, so the normalization factor is $1 - C = 0.0199$. The fused belief for class C_1 becomes $b_{C_1} = (0.99 \times 0 + 0.99 \times 0.01 + 0 \times 0.01)/0.0199 \approx 0.4975$. Similarly, $b_{C_2} \approx 0.4975$, and $b_{C_3} = 0$. The fused uncertainty is $u = (0.01 \times 0.01)/0.0199 \approx 0.005$. This result assigns nearly equal belief to C_1 and C_2 with very low uncertainty, which is counter-intuitive, since the fused opinion expresses high confidence despite the obvious conflict between the two experts. The issue arises because the conflicting mass (the product of conflicting beliefs) is entirely discarded by Dempster's rule rather than being redistributed to reflect the lack of consensus. To solve this drawback under a multi-expert framework, we introduce a Jensen-Shannon divergence-based weighted average mass [64] to pre-process the opinions before applying Dempster's rule. In this way, the final fusion considers not only belief and uncertainty, but also inter-expert agreement and intra-expert information volume.

Specifically, we first calculate the distance matrix $D = (d_{ij})_{K \times K}$ based on Jensen-Shannon divergence to measure the conflict degree of the opinions as

$$d_{ij} = \frac{1}{2}\left[S\left(\mathcal{M}^i, \frac{\mathcal{M}^i + \mathcal{M}^j}{2}\right) + S\left(\mathcal{M}^j, \frac{\mathcal{M}^i + \mathcal{M}^j}{2}\right)\right],$$

where $S(\mathcal{M}^i, \mathcal{M}^j) = \sum_{m=1}^M b_m^i \log(b_m^i/b_m^j) + u^i \log(u^i/u^j)$. Obviously, $d_{ii} = 0$. Moreover, we quantify the information volume $\{v_i\}_{i=1}^K$ of the expert opinions using mass entropy, given by $v_i = -\sum_{m=1}^M b_m^i \log(b_m^i) - u^i \log(u^i)$. A small v_i indicates that most of the mass is concentrated on either a single class or the uncertainty term. Under noisy labels, the former may reflect memorization of corrupted supervision, whereas the latter suggests insufficient class-specific evidence. Compared with these two cases, a larger v_i indicates that the opinion retains belief in alternative class hypotheses together with a certain degree of uncertainty, instead of collapsing onto a single component. Such an opinion is useful under noisy supervision because it reduces the risk that a potentially corrupted label dominates the prediction. The benefit of a larger v_i therefore lies in avoiding overconfident or uninformative collapse rather than directly providing stronger class discrimination. However, mass entropy alone is still insufficient for assessing an expert, since a nearly uniform opinion has high entropy but little discriminative power. Hence, we also use the total Jensen-Shannon divergence to measure the agreement between an expert and the other experts. A small total divergence indicates that the opinion is supported by similar assessments from the other experts, whereas a large divergence indicates an isolated or conflicting opinion. We use these two quantities together so that an important expert should both retain sufficient information about alternative predictions and receive support from the other experts. Based on these considerations, we define the importance degree of the ith expert as

$$c_i = \frac{\tilde{c}_i}{\sum_{k=1}^K \tilde{c}_k} \quad \text{with} \quad \tilde{c}_i = v_i \left(\sum_{j=1, j\neq i}^K d_{ij}\right)^{-1}.$$

On account of the importance degrees $\{c_i\}_{i=1}^K$, we introduce the weighted average opinion as

$$\overline{\mathcal{M}} = \{\bar{b}_1, \dots, \bar{b}_M, \bar{u}\} = \sum_{i=1}^{K} c_i \mathcal{M}^i.$$

Lastly, the weighted average opinion $\overline{\mathcal{M}}$ is further fused via Dempster's combination rule by $K-1$ times to achieve the fused multinomial opinion at the decision level, i.e.,

$$\mathcal{M}^{\text{Fusion}} = \{b_1^{\text{D}}, \dots, b_M^{\text{D}}, u^{\text{D}}\} = \underbrace{\overline{\mathcal{M}} \oplus \cdots \oplus \overline{\mathcal{M}}}_{K-1 \text{ times}}. \tag{4}$$

Through the pre-processing of the multinomial opinion from each expert, we can effectively resist the counter-intuitive phenomena by incorporating the measure of importance degree. Moreover, Dempster's rule is adopted naturally, which preserves the desirable properties of the fusion rule, part of which can be derived as follows.

Theorem 1. *Let $\mathcal{M} = \{b_1, \dots, b_M, u\} = \overline{\mathcal{M}} \oplus \overline{\mathcal{M}}$ be the fused multinomial opinion from $\overline{\mathcal{M}}$ by Dempster's combination rule and t be the index of the ground-truth class. If $\bar{b}_t \geq 1 - 2\bar{u}$ or $\bar{b}_t = \bar{b}_{\max} \triangleq \max\{\bar{b}_1, \dots, \bar{b}_M\}$, then we have $b_t \geq \bar{b}_t$.*

Proof. From the expression of $\overline{\mathcal{M}} = \sum_{i=1}^K c_i \mathcal{M}^i$, the normalization property still holds, i.e., $\sum_{m=1}^M \bar{b}_m + \bar{u} = 1$. Moreover, according to Definition 1, we have

$$b_t = \frac{\bar{b}_t^2 + 2\bar{b}_t\bar{u}}{1 - \sum_{i\neq j} \bar{b}_i\bar{b}_j} = \frac{\bar{b}_t^2 + 2\bar{b}_t\bar{u}}{\sum_{m=1}^M \bar{b}_m^2 + 2\bar{u} - \bar{u}^2} \geq \frac{\bar{b}_t^2 + 2\bar{b}_t\bar{u}}{(\sum_{m=1}^M \bar{b}_m)^2 + 2\bar{u} - \bar{u}^2} = \bar{b}_t(\bar{b}_t + 2\bar{u}).$$

If $\bar{b}_t \geq 1 - 2\bar{u}$, we obtain $b_t \geq \bar{b}_t$. In addition,

$$b_t = \frac{\bar{b}_t^2 + 2\bar{b}_t\bar{u}}{\sum_{m=1}^M \bar{b}_m^2 + 2\bar{u} - \bar{u}^2} \geq \frac{\bar{b}_t(\bar{b}_t + 2\bar{u})}{\bar{b}_{\max}(1-\bar{u}) + 2\bar{u}} \geq \frac{\bar{b}_t(\bar{b}_t + 2\bar{u})}{\bar{b}_{\max} + 2\bar{u}}.$$

On the condition that $\bar{b}_t = \bar{b}_{\max}$, we can also derive $b_t \geq \bar{b}_t$, which completes the proof. □

Theorem 2. *Let $\mathcal{M} = \{b_1, \dots, b_M, u\} = \mathcal{M}^i \oplus \mathcal{M}^j$ be the Dempster's fused opinion from any two multinomial opinions and t be the index of the ground-truth class. Denote $b_t^{\max} = \max\{b_t^i, b_t^j\}$, $b_t^{\min} = \min\{b_t^i, b_t^j\}$, and $b_{\max}^i = \max\{b_1^i, \dots, b_M^i\}$. If*

$$\max\left\{\frac{u^i + u^j}{1 + u^j - u^i u^j}, \frac{u^j + u^i}{1 + u^i - u^i u^j}\right\} \geq 1 \text{ or } b_t^{\max} \geq \min\{b_{\max}^i, b_{\max}^j\},$$

then we have $b_t \geq b_t^{\min}$.

Proof. Similar to the proof of Theorem 1, we have

$$b_t = \frac{b_t^i b_t^j + b_t^i u^j + b_t^j u^i}{\sum_{m=1}^M b_m^i b_m^j + u^i + u^j - u^i u^j} \geq \frac{b_t^{\min}(b_t^{\max} + u^j + u^i)}{\sum_{m=1}^M b_m^i + u^i + u^j - u^i u^j} = \frac{b_t^{\min}(b_t^{\max} + u^j + u^i)}{1 + u^j - u^i u^j} \geq \frac{b_t^{\min}(u^j + u^i)}{1 + u^j - u^i u^j}.$$

On the condition that $(u^j + u^i)/(1 + u^j - u^i u^j) \geq 1$, we have $b_t / b_t^{\min} \geq 1$. In addition, similar to [65]

$$b_t = \frac{b_t^i b_t^j + b_t^i u^j + b_t^j u^i}{\sum_{m=1}^M b_m^i b_m^j + u^i + u^j - u^i u^j} \geq \frac{b_t^{\min}(b_t^{\max} + u^j + u^i)}{b_{\max}^i(1-u^j) + u^i + u^j - u^i u^j} \geq \frac{b_t^{\min}(b_t^{\max} + u^j + u^i)}{b_{\max}^i + u^i + u^j}.$$

On the condition that $b_t^{\max} \geq b_{\max}^i$, we can also obtain that $b_t / b_t^{\min} \geq 1$. The above conclusions are also true when the index $i(/j)$ is replaced by $j(/i)$ in the inequality, which completes the proof. □

Theorem 3. *When Dempster's combination rule is performed on $\overline{\mathcal{M}}$, the fused uncertainty mass can be reduced. Furthermore, a smaller uncertainty mass $\bar{u}$ would lead to a greater degree of uncertainty reduction.*

Proof. By Definition 1, the fused uncertainty mass

$$u = \frac{\bar{u}^2}{\sum_{i=1}^M \bar{b}_i^2 + 2\bar{u} - \bar{u}^2} \leq \frac{\bar{u}}{2 - \bar{u}},$$

indicating that $u \leq \bar{u}$. Moreover, the inequality $u/\bar{u} \leq 1/(2-\bar{u})$ implies that the relative reduction degree of u is positively correlated with the descending extent of $\bar{u}$. □

Theorems 1–3 analyze how the Dempster combination stage affects the belief assigned to the ground-truth class and the epistemic uncertainty of the resulting opinion. Theorem 1 analyzes the basic self-combination $\overline{\mathcal{M}} \oplus \overline{\mathcal{M}}$ in Eq. (4). It identifies two sufficient conditions under which this operation preserves or increases the belief assigned to the ground-truth class. Theorem 2 considers a general binary combination and establishes conditions under which fusion does not reduce the ground-truth belief below the lower of the two input beliefs. This result applies to the sequential computation of Eq. (4), in which an intermediate fused opinion is combined with another copy of $\overline{\mathcal{M}}$. Theorem 3 focuses on the uncertainty component and shows that self-combination reduces the uncertainty mass. It therefore explains how the Dempster combination stage concentrates the weighted opinion by transferring mass

from uncertainty to class-specific beliefs. While some conclusions are consistent with those reported in [65], our analysis derives additional sufficient conditions and covers the self-combination operation used in Eq. (4). Together, these theorems characterize the evolution of class belief and uncertainty during the combination process in EvGNN and provide theoretical support for the soundness of our decision fusion design. In particular, they show that applying Dempster's rule can preserve or strengthen a sufficiently supported ground-truth belief while reducing the remaining uncertainty mass. Nevertheless, they do not guarantee that the ground-truth class always receives the largest fused belief. We therefore introduce the following theorem to derive an upper bound on the probability that the fused belief assigned to the observed noisy class is no smaller than that assigned to the ground-truth class.

Theorem 4. *Let t and $\tilde{t}$ denote the ground-truth class and observed noisy class, respectively. Define $X_i = \mathbb{I}_{\{b_{\tilde{t}}^i > b_t^i\}}$ and $p_i^{\mathrm{mis}} = \mathbb{P}(X_i = 1)$ for the ith expert. Assume that the experts are not systematically misled by the same observed label, in the sense that $\mathbb{P}\big(X_i = 1 \mid X_j = 1 \text{ for all } j \in \mathcal{J}\big) \le \lambda p_i^{\mathrm{mis}}$ for any i and any $\mathcal{J} \subseteq \{1, \dots, K\} \setminus \{i\}$, where $\lambda \ge 1$ measures the dependence among expert errors. When $\sum_{j=1}^K c_j |b_{\tilde{t}}^j - b_t^j| > 0$, denote $\omega_i = \frac{c_i |b_{\tilde{t}}^i - b_t^i|}{\sum_{j=1}^K c_j |b_{\tilde{t}}^j - b_t^j|}$, $\bar{p}_{\mathrm{mis}} = \sum_{i=1}^K \omega_i p_i^{\mathrm{mis}}$, $\omega_{\max} = \max_i \omega_i$. If $\lambda \bar{p}_{\mathrm{mis}} < 1/2$, then*

$$\mathbb{P}\big(b_{\tilde{t}}^{\mathrm{D}} \ge b_t^{\mathrm{D}}\big) \le \exp\left\{ -\frac{1}{\omega_{\max}} \left[\frac{1}{2} \log \frac{1}{2\lambda \bar{p}_{\mathrm{mis}}} - \frac{1}{2} + \lambda \bar{p}_{\mathrm{mis}} \right] \right\}.$$

Proof. For any nonempty subset $\mathcal{J} = \{j_1, \dots, j_r\}$, the chain rule gives

$$\mathbb{P}(X_j = 1,\ j \in \mathcal{J}) = p_{j_1}^{\mathrm{mis}} \prod_{l=2}^{r} \mathbb{P}\Big(X_{j_l} = 1 \mid X_{j_1} = \cdots = X_{j_{l-1}} = 1 \Big) \le \lambda^{r-1} \prod_{j \in \mathcal{J}} p_j^{\mathrm{mis}}.$$

For any $\theta > 0$, expanding the moment-generating function over all expert subsets yields

$$\mathbb{E} \exp\left(\theta \sum_{i=1}^K \omega_i X_i \right) \le 1 + \frac{1}{\lambda} \left\{ \prod_{i=1}^K \left[1 + \lambda p_i^{\mathrm{mis}} \left(e^{\theta \omega_i} - 1 \right) \right] - 1 \right\} \le \exp\left[\sum_{i=1}^K \lambda p_i^{\mathrm{mis}} \left(e^{\theta \omega_i} - 1 \right) \right].$$

Since $0 \le \omega_i \le \omega_{\max}$ and $e^{\theta x} - 1$ is convex on $[0, \omega_{\max}]$, we have $e^{\theta \omega_i} - 1 \le \frac{\omega_i}{\omega_{\max}} \left(e^{\theta \omega_{\max}} - 1 \right)$. Therefore,

$$\mathbb{E} \exp\left(\theta \sum_{i=1}^K \omega_i X_i \right) \le \exp\left[\frac{\lambda \bar{p}_{\mathrm{mis}}}{\omega_{\max}} \left(e^{\theta \omega_{\max}} - 1 \right) \right].$$

By Markov's inequality,

$$\mathbb{P}\left(\sum_{i=1}^K \omega_i X_i \ge \frac{1}{2} \right) \le \exp\left(-\frac{\theta}{2} \right) \mathbb{E}\left[\exp\left(\theta \sum_{i=1}^K \omega_i X_i \right) \right] \le \exp\left[-\frac{\theta}{2} + \frac{\lambda \bar{p}_{\mathrm{mis}}}{\omega_{\max}} \left(e^{\theta \omega_{\max}} - 1 \right) \right].$$

When $\lambda \bar{p}_{\mathrm{mis}} < 1/2$, the right-hand side is minimized at $\theta = \omega_{\max}^{-1} \log(1/(2\lambda \bar{p}_{\mathrm{mis}}))$, which gives the exponential bound.

It remains to connect the event $\left\{ \sum_{i=1}^K \omega_i X_i \ge \frac{1}{2} \right\}$ to the fused belief. From the definition of ω_i, we have

$$\frac{\bar{b}_{\tilde{t}} - \bar{b}_t}{\sum_{i=1}^K c_i |b_{\tilde{t}}^i - b_t^i|} = 2 \sum_{i=1}^K \omega_i X_i - 1.$$

Under the Dempster's rule in Eq. (4), we can obtain that

$$b_{\tilde{t}}^{\mathrm{D}} - b_t^{\mathrm{D}} = \frac{(\bar{b}_{\tilde{t}} + \bar{u})^{K-1} - (\bar{b}_t + \bar{u})^{K-1}}{\sum_{m=1}^M \left[(\bar{b}_m + \bar{u})^{K-1} - \bar{u}^{K-1} \right] + \bar{u}^{K-1}}.$$

The denominator is positive, and x^{K-1} is increasing on $[0, 1]$. Hence, $b_{\tilde{t}}^{\mathrm{D}} \ge b_t^{\mathrm{D}}$ holds if and only if $\sum_i \omega_i X_i \ge 1/2$. □

Theorem 4 characterizes how expert diversity and adaptive opinion weighting jointly control the influence of misleading predictions in EvGNN. Different parameter initializations and optimization trajectories encourage heterogeneous prediction errors; when these errors are less strongly aligned, the dependence factor λ becomes smaller, leading to a tighter bound. Moreover, the weighting scheme based on JS divergence and information volume assigns less importance to experts that strongly disagree with the others or provide less informative evidence. If such an expert is likely to favor the observed noisy class, reducing its importance limits the term $\omega_i p_i^{\mathrm{mis}}$ in $\bar{p}_{\mathrm{mis}}$ and prevents it from exerting excessive influence through $\omega_{\max}$. Since the condition $\lambda \bar{p}_{\mathrm{mis}} < 1/2$ constrains only the collective weighted misleading influence, EvGNN can tolerate individually unreliable experts as long as their contributions are appropriately controlled. When useful contributions are evenly distributed across experts such that $\omega_{\max} = 1/K$, the upper bound further decreases exponentially with K. Thus, the theorem establishes sufficient conditions under which the fused decision remains robust despite the presence of unreliable experts.

To complement the robustness provided by decision fusion, we impose evidence-based supervision on the fused opinion, allowing it to be jointly optimized with the individual experts rather than treated merely as a post-processing result. Based on the fused multinomial opinion $\mathcal{M}^{\mathrm{Fusion}}$ in Eq. (4), the fused evidence from K experts and the fused-version Dirichlet parameters for any node are derived as

$$e_m^{\mathrm{D}} = b_m^{\mathrm{D}} S^{\mathrm{D}},\ S^{\mathrm{D}} = \frac{M}{u^{\mathrm{D}}},\ \alpha_m^{\mathrm{D}} = e_m^{\mathrm{D}} + 1,\ \text{for } m = 1, \dots, M.$$

Similar to Eq. (2), we induce the fused evidence-based supervised loss with respect to $\boldsymbol{\alpha}^{\mathrm{D}} = (\alpha_1^{\mathrm{D}}, \ldots, \alpha_M^{\mathrm{D}})^\top$ as

$$\mathcal{L}_{evi_fus} = \sum_{s=1}^{L} \sum_{m=1}^{M} y_{sm}(\psi(S_s^{\mathrm{D}}) - \psi(\alpha_{sm}^{\mathrm{D}})), \tag{5}$$

where $S_s^{\mathrm{D}} = M/u_s^{\mathrm{D}}$; $\boldsymbol{\alpha}_s^{\mathrm{D}} = (\alpha_{s1}^{\mathrm{D}}, \ldots, \alpha_{sM}^{\mathrm{D}})^\top$ and u_s^{D} are the fused Dirichlet parameters and uncertainty for the sth node in $\mathcal{V}_{\mathrm{L}}$, respectively. Accordingly, we incorporate the measure of uncertainty by combining the evidence-based losses for all experts with the fusion version at the decision level as

$$\mathcal{L}_{evi_total} = \mathcal{L}_{evi} + \mathcal{L}_{evi_fus}. \tag{6}$$

The joint optimization provides an optimization-level complement to model robustness. In this way, EvGNN couples evidence-based supervision for alleviating the adverse effects of noisy labels with conflict-aware opinion fusion for mitigating counter-intuitive fusion outcomes, thereby producing more reliable and robust predictions under label noise.

3.3. Representation fusion of multiple experts

Based on the dependent topological structure among the nodes, semi-supervised learning on graphs can implicitly instruct the representation learning of unlabeled nodes by supervised loss derived from labeled nodes through message-passing. This however fails to sufficiently leverage the information from unlabeled nodes, which may result in poor generalization ability, especially when the labels are sparse and noisy. Hence, besides the information fusion at the decision level, we also reinforce the representation-level fusion, which can be utilized coupled with consistency learning to directly enhance the attention to unlabeled nodes, leading to higher model robustness.

On the basis of the GCN-based embeddings $\{\mathbf{Z}_i \in \mathbb{R}^{N\times d'}\}_{i=1}^{K}$ from K experts, the goal of representation-level fusion is to integrate the expert-specific embeddings into a compact representation. As a flexible tool, the multi-layer perceptron (MLP) is able to generate effective or expressive coding. To achieve semantically rich, expressive, and compact representations $\mathbf{Z}^{\mathrm{Fusion}}$, we concatenate the expert-specific embeddings and perform an MLP-based deep fusion on them, which can be formulated as

$$\mathbf{Z}^{\mathrm{Fusion}} = f([\mathbf{Z}^1 || \cdots || \mathbf{Z}^K]), \tag{7}$$

where $\cdot||\cdot$ denotes the concatenation operation and f is the fusion rule. The concatenation operation explicitly preserves the identity of each expert, allowing the fusion network to directly access individual expert representations and learn relationships among different experts. It provides greater flexibility for capturing distinctive characteristics and fine-grained cross-expert interactions that may be informative for representation fusion. In the fusion rule f, we employ two residual MLP blocks with complementary bottleneck structures to refine the concatenated expert representations. Let $\mathbf{Z}^{(0)} = [\mathbf{Z}^1 || \cdots || \mathbf{Z}^K] \in \mathbb{R}^{N\times Kd'}$ denote the concatenated representation. The two residual blocks are defined as

$$\mathbf{Z}^{(1)} = \mathbf{Z}^{(0)} + \mathrm{MLP}_{2Kd'\to Kd'}(\sigma(\mathrm{MLP}_{Kd'\to 2Kd'}(\mathbf{Z}^{(0)}))), \mathbf{Z}^{\mathrm{Fusion}} = \mathbf{Z}^{(1)} + \mathrm{MLP}_{Kd'/2\to Kd'}(\sigma(\mathrm{MLP}_{Kd'\to Kd'/2}(\mathbf{Z}^{(1)}))),$$

where $\mathrm{MLP}_{a\to b}$ represents a single fully connected layer that maps an a-dimensional input to a b-dimensional output, and $\sigma(\cdot)$ denotes the activation function, e.g., ReLU and Softplus. The first block adopts an inverted bottleneck design, i.e., $Kd' \to 2Kd' \to Kd'$, while the second block adopts a standard bottleneck design, i.e., $Kd' \to Kd'/2 \to Kd'$. Residual connections [66] are introduced to stabilize optimization and each residual MLP block contains two fully connected layers. The expansion-reduction block is placed first because the concatenated representation contains heterogeneous information from different experts. Expanding the dimension from Kd' to $2Kd'$ provides a larger intermediate space for modeling nonlinear cross-expert interactions, while the following reduction maps the enriched features back to the original residual dimension. The second block follows a reduction-expansion design that compresses $\mathbf{Z}^{(1)}$ from Kd' to $Kd'/2$ and then projects it back to the original dimension. The compact intermediate representation encourages the model to retain discriminative information while suppressing redundant or noisy expert-specific features. In this way, the first residual block focuses on modeling cross-expert interactions and enriching the representation, whereas the second residual block focuses on compact feature extraction and redundancy suppression.

3.4. Consistency and pseudo-label learning module

Consistency learning [67] can boost the exploration and utilization of information from a large number of unlabeled nodes on the graph and prevent GNNs from overfitting to noisy labels, whereas pseudo-labeling [68] can provide more confident supervision to resist the label scarcity problem. To benefit from both techniques, we focus on the consistency learning between decision- and representation-level fusions from Sections 3.2 and 3.3 and the pseudo-labeling based on the fused evidence from Section 3.2.

For consistency learning, we compare the evidences based on the decision and representation fusions under the subjective logic framework. The purpose is to make the decision-level and representation-level evidence mutually regularize each other, which enables nodes to provide an additional uncertainty-aware training signal under label scarcity. Specifically, we first produce the Dirichlet parameters from the fused representation as

$$\boldsymbol{\alpha}_s^{\mathrm{R}} = \mathbf{e}_s^{\mathrm{R}} + 1, \text{ for } s = 1, \ldots, N, \tag{8}$$

where $\mathbf{e}_s^{\mathrm{R}}$ is the class evidence generated from the fused representation corresponding to the sth node in $\mathcal{V}$, i.e., feed the sth row of $\mathbf{Z}^{\mathrm{Fusion}}$ into an MLP to project the dimensionality to the class number M. Then we minimize the KL divergence between the Dirichlet distributions parameterized by $\boldsymbol{\alpha}_s^{\mathrm{D}}$ in Eq. (5) and $\boldsymbol{\alpha}_s^{\mathrm{R}}$ in Eq. (8), i.e., the consistency learning loss is defined as

$$\mathcal{L}_{con} = \sum_{s=1}^{N} \mathrm{KL}(Dir(\mathbf{p}_s^{\mathrm{R}}|\boldsymbol{\alpha}_{\mathbf{s}}^{\mathrm{R}})|\mathbf{Dir}(\mathbf{p}_{\mathbf{s}}^{\mathrm{D}}|\boldsymbol{\alpha}_{\mathbf{s}}^{\mathrm{D}})). \tag{9}$$

In this way, we leverage all nodes to enhance the self-reinforcement consistency learning without querying label information, and also limit the effect of label noise by the subjective logic framework, which can effectively improve the robustness of the proposed framework.

For pseudo-labeling, we leverage the decision-level fused evidences at the tth epoch to predict the node labels for all nodes in $\mathcal{V}$, including both labeled and unlabeled nodes. We regard predictions with high confidence as reliable pseudo labels, which can subsequently guide the training at the $(t+1)$th epoch, thereby providing auxiliary supervision. Formally, we first predict the labels of all nodes in $\mathcal{V}$ by

$$\hat{m}_s^{(t)} = \mathrm{argmax}_m \ \mathrm{softmax}(\{e_{sm}^{\mathrm{D}(t)}\}_{m=1}^{M}) \triangleq \mathrm{argmax}_m \mathbf{s}_s^{\mathrm{D}(t)}, \ \text{for } s = 1, \ldots, N, \tag{10}$$

where $\hat{m}_s^{(t)}$ is the predicted label for the sth node in $\mathcal{V}$, $\{e_{sm}^{\mathrm{D}(t)}\}_{m=1}^{M}$ are the evidences obtained from the tth epoch, and $\mathbf{s}_s^{\mathrm{D}(t)}$ is the score vector of the evidence. Then, we preserve the pseudo labels whose largest evidence falls above a predefined threshold τ to supervise model learning. The corresponding pseudo-labeling loss is defined as

$$\mathcal{L}_{pse} = -\sum_{s=1}^{N}\sum_{m=1}^{M} \mathbb{I}_{\{\max \mathbf{s}_s^{\mathrm{D}(t)} \geq \tau\}} \hat{y}_{sm}^{(t)} \log(\mathbf{s}_s^{\mathrm{D}(t+1)}), \tag{11}$$

where $\hat{\mathbf{y}}_s^{(t)} = (\hat{y}_{s1}^{(t)}, \ldots, \hat{y}_{sM}^{(t)})^{\top}$ is the one-hot pseudo label according to $\hat{m}_s^{(t)}$ for the sth node in $\mathcal{V}$, i.e., $\hat{y}_{s\hat{m}_s}^{(t)} = 1$, and $\mathbb{I}_{(\cdot)}$ is the indicator function that returns 1 if the condition is satisfied and returns 0 otherwise. Note that $\mathcal{L}_{pse}$ uses pseudo labels generated from the decision-level fused evidence rather than directly from the observed labels. Therefore, pseudo-labeling does not explicitly rely on the original noisy annotations. For labeled nodes, the evidence-based supervised loss still uses the observed labels, but its evidential formulation explicitly models uncertainty, which helps avoid overconfident fitting to potentially unreliable annotations. The pseudo-labeling loss further complements the uncertainty-aware supervised loss by providing an additional confidence-based supervision signal from high-confidence fused predictions. When the pseudo label disagrees with the observed label, it can further mitigate the influence of potentially corrupted labels. In the experiments, we set the threshold τ as 0.95 to ensure the high confidence of the pseudo labels.

Algorithm 1 The pseudo-code of EvGNNEvGNN.

Input: Graph $\mathcal{G} = \{\mathcal{V} = \mathcal{V}_{\mathrm{L}} \cup \mathcal{V}_{\mathrm{U}}, \mathcal{E}, \mathbf{X}, \mathbf{Y}_{\mathrm{L}}\}$; Class number M; Expert number K; Maximum iterations $T_{\max}$;
Output: Classification prediction $\hat{\mathbf{Y}}_{\mathrm{U}}$;

```
Initialize the parameters in GCN for K experts;
for t = 1 to T_max do
    Calculate evidence {e_m^i}_{m=1}^M and uncertainty u^i for i = 1, ..., K by Eq. (1);
    Update multinomial opinions {M^1, ..., M^K} and generate the fused opinion M^Fusion by Section 3.2;
    Calculate the loss L_evi_total in Eq. (6);
    Update the fused representation in Eq. (7);
    Update the evidence based on fused representation in Eq. (8) and calculate the loss L_con in Eq. (9);
    Update the pseudo-labels in Eq. (10) and calculate the loss L_pse in Eq. (11);
    Conduct the backpropagation and update the whole network by minimizing the total loss L in Eq. (12);
end for
Employ the form of Dirichlet mean, i.e., p̂ = α^D / S^D, to predict the class probability assignment for each unlabeled node in V_U;
return Ŷ_U;
```

3.5. Joint optimization module for robust GNN

In this paper, our primary focus is on robust learning on graphs with noisy and sparse labels. To this end, we conduct evidence-based uncertainty quantification, decision- and representation-level fusions, consistency learning, and pseudo-labeling, which effectively alleviates the negative impact of noisy labels and improves the utilization of unlabeled information against label sparsity.

Formally, we unite the evidence-based loss in Eq. (6), consistency learning loss in Eq. (9), and pseudo-labeling loss in Eq. (11) to optimize EvGNN, i.e., the total loss is

$$\mathcal{L} = \mathcal{L}_{evi_total} + \eta_1 \mathcal{L}_{con} + \eta_2 \mathcal{L}_{pse}, \tag{12}$$

where η_1 and η_2 are hyper-parameters to control the relative importance of each loss component.

Finally, when the model converges, for each unlabeled node in $\mathcal{V}_U$, we employ the form of Dirichlet mean, i.e., $\hat{\mathbf{p}} = \boldsymbol{\alpha}^D / S^D$, to predict the class probability assignment in an evidence-ensemble view. The training process of our proposed EvGNN is summarized in Algorithm 1.

Computational Complexity Analysis. Given a graph with N nodes and E edges, and the dimensions of C-layer GCN are $d_1, d_2, \ldots, d_C$. For K experts, the time complexity of producing representations and evidences from the GCN encoder is $O(KE\sum_{i=1}^{C} d_i d_{i-1} + KNMd_C)$. We perform decision fusion in $O(K^2M + M^2K)$ and representation fusion in $O(NK^2d_C^2 + NKMd_C)$. Moreover, the complexity of the total loss is $O(KLM + NM)$. Hence, the total computational complexity of EvGNN is $O(KE\sum_{i=1}^{C} d_i d_{i-1} + KNMd_C + NK^2d_C^2)$.

4. Experiments

4.1. Experimental setup

Datasets. We evaluate the experiments on six widely used benchmark datasets[1], including two author networks: Coauthor CS and DBLP, two co-purchase networks: Amazon Computers and Amazon Photo, and two citation networks: Pubmed and Citeseer. Following [16], 80% of the nodes are set as the test set and 10% as the validation set. As for the training set, we randomly select 10% (i.e., 1% of all nodes) nodes in the remaining 10% node set for the large datasets Coauthor CS, Amazon Photo, Amazon Computers, Pubmed, and DBLP and select 50% (i.e., 5% of all nodes) of the remaining nodes for Citeseer as the labeled nodes. It is worth noting that in transductive graph learning, all nodes in the full graph are visible during model training, while the label information of the test and validation sets is not accessible. Therefore, in our setting, the amount of labeled nodes is very limited compared to the number of nodes involved in training, i.e., only roughly 1% or 5% of the entire dataset. Since all labels in the datasets are clean, following [8,11,16], we pollute the labels in training set with two types of label noises, i.e., *uniform noise*: labels flip to any other class with probability $p/(M-1)$; *pair noise*: labels only flip to their closest class with probability p.

Baselines. To illustrate the superiority of our EvGNN, we compare its performance with some classic GNN methods, i.e., the popular GNNs: GCN [1], GIN [36]; and other methods designed for noisy labels: Forward [11], Coteaching + [8], D-GNN [69], CP [70], NRGNN [16], RTGNN [17], CGNN [18], and CR-GNN [19]; multi-expert- or Transformer-based methods: RNCGLN [71], MEGC [72], KDEM and PPEM [73]. Except for GNNs, all the other methods including EvGNN adopt GCN as the backbone for a fair comparison. For NRGNN and RTGNN, we have conducted a rerun of their experiments to verify the results. For CGNN and CR-GNN, we follow the code provided by NoisyGL [45] and also perform hyper-parameter optimization to ensure the best performance. For RNCGLN [71], MEGC [72], and KDEM and PPEM [73], we either reproduce the methods or use their publicly available implementations and evaluate their performance under our experimental settings.

Implementation details. To optimize all models, we leverage the Adam optimizer with a learning rate of 0.0001 for Coauthor CS and Pubmed, and 0.001 for other datasets. For our EvGNN, we set the expert number K as 3 to balance performance and efficiency, and the noise rate p as 20%. We separately vary the range of the hyper-parameter η_1 from 0.001 to 10 and η_2 from 0.01 to 10. The pseudo-label threshold τ is fixed at 0.95. The maximum number of training epochs is 1000. Following [16], each method is replicated for 5 runs to report the mean accuracy and standard deviation on the test set for evaluation.

4.2. Overall comparison

In this section, we evaluate the performance of our EvGNN and all baseline methods for node classification on graphs. The experimental results on six benchmarks under two types of label noises are shown in Table 1. Based on the quantitative results, we can observe the following points:

- The performance of classic GNN methods (GCN and GIN) is inferior to most approaches designed for noisy labels, which suggests that these GNNs may be ineffective in resisting the impact of noisy labels, and are prone to overfitting noisy features, leading to poor generalization. Moreover, the limited labels also restrict the model from obtaining sufficient supervisory signals, resulting in sub-optimal performance.
- Compared with Forward and Coteaching +, graph-specific methods for noisy-label learning, including D-GNN, CP, NRGNN, RTGNN, CGNN, CR-GNN, RNCGLN, and MEGC, generally exhibit superior performance. It indicates that incorporating graph structures and semantics is important for mitigating the effects of noisy supervision on graph-related tasks.
- Among all baseline methods, several state-of-the-art approaches exhibit a phenomenon of flourishing diversity, each having its own strengths across different datasets. NRGNN and RTGNN effectively propagate messages between labeled and unlabeled nodes through connections, while CGNN and CR-GNN leverage contrastive learning to better integrate neighbor information for learning more intrinsic graph-semantic node representations, mitigating to varying degrees the side effects of noisy labels. RNCGLN combines local graph contrastive learning and pseudo-graph refinement with Transformer-style attention to model both local structural information and global node interactions. MEGC employs collaborative experts for graph-structure governance and label-noise governance, separately addressing structural perturbations and unreliable supervision. KDEM and PPEM integrate

[1] https://github.com/shchur/gnn-benchmark/raw/master/data/npz/

Table 1
Performance on six benchmark datasets (mean ± std). The best results among all the methods and all the baselines are highlighted with **bold** and underline, respectively.

Method	Noise	Coauthor CS	Amazon Photo	Amazon Computers	Pubmed	DBLP	Citeseer
GCN	Uniform	80.3 ± 1.4	82.2 ± 0.9	76.9 ± 1.2	77.3 ± 0.9	71.0 ± 1.5	64.9 ± 1.7
	Pair	79.5 ± 1.1	80.9 ± 1.2	76.3 ± 1.6	78.0 ± 0.4	72.5 ± 1.2	60.3 ± 1.0
GIN	Uniform	80.5 ± 1.2	82.1 ± 0.4	76.4 ± 0.7	77.4 ± 0.5	72.4 ± 0.7	65.7 ± 2.1
	Pair	79.3 ± 1.7	81.2 ± 1.5	76.8 ± 1.2	78.1 ± 0.6	73.4 ± 2.1	61.6 ± 1.0
Forward	Uniform	81.3 ± 0.4	79.8 ± 0.6	78.4 ± 0.4	77.5 ± 0.4	73.1 ± 0.3	65.0 ± 1.5
	Pair	80.5 ± 0.8	78.7 ± 0.3	77.3 ± 0.6	79.6 ± 0.2	74.4 ± 0.5	61.6 ± 0.4
Coteaching +	Uniform	80.7 ± 1.4	78.5 ± 0.6	70.3 ± 3.6	78.6 ± 0.4	73.5 ± 1.3	66.4 ± 1.3
	Pair	77.6 ± 3.3	75.5 ± 1.8	60.9 ± 3.9	78.5 ± 0.1	72.7 ± 1.2	65.1 ± 2.1
D-GNN	Uniform	81.9 ± 1.1	84.5 ± 3.4	79.5 ± 0.8	77.6 ± 0.3	72.8 ± 1.2	64.9 ± 1.3
	Pair	79.9 ± 0.7	83.2 ± 2.2	79.8 ± 1.2	79.4 ± 0.4	75.4 ± 0.9	62.3 ± 1.2
CP	Uniform	82.5 ± 1.2	83.7 ± 0.7	80.6 ± 1.3	78.6 ± 0.3	74.2 ± 0.5	66.0 ± 1.6
	Pair	80.2 ± 1.0	83.1 ± 1.1	80.1 ± 2.4	77.9 ± 0.3	73.6 ± 1.0	62.0 ± 1.0
NRGNN	Uniform	83.2 ± 0.5	83.7 ± 3.9	80.3 ± 3.2	79.0 ± 1.6	79.5 ± 1.5	70.1 ± 1.7
	Pair	83.7 ± 0.9	83.5 ± 3.6	79.6 ± 3.4	79.2 ± 0.7	79.3 ± 0.9	67.8 ± 3.0
RTGNN	Uniform	86.9 ± 1.6	84.8 ± 3.3	81.0 ± 2.3	79.8 ± 1.3	79.0 ± 1.1	68.2 ± 3.8
	Pair	83.8 ± 2.1	84.2 ± 2.7	80.3 ± 2.0	80.4 ± 1.6	78.4 ± 2.6	67.0 ± 2.8
CGNN	Uniform	84.1 ± 0.4	85.3 ± 0.9	81.8 ± 0.3	78.1 ± 0.4	76.8 ± 0.5	69.7 ± 1.3
	Pair	81.0 ± 1.1	85.1 ± 0.7	81.6 ± 0.2	78.6 ± 0.4	77.5 ± 0.4	66.0 ± 1.7
CR-GNN	Uniform	82.9 ± 2.3	81.5 ± 4.6	79.1 ± 4.2	80.3 ± 1.1	79.2 ± 1.3	69.3 ± 2.1
	Pair	81.7 ± 3.1	78.1 ± 5.2	78.6 ± 3.8	80.1 ± 1.2	78.9 ± 1.7	68.6 ± 1.9
RNCGLN	Uniform	84.4 ± 2.1	71.8 ± 3.9	70.4 ± 1.5	76.9 ± 1.4	75.5 ± 2.0	71.4 ± 1.4
	Pair	83.4 ± 2.7	74.5 ± 3.2	68.4 ± 2.3	73.5 ± 2.7	73.1 ± 3.0	**71.0 ± 1.5**
MEGC	Uniform	84.4 ± 1.1	80.8 ± 4.9	78.1 ± 1.8	77.3 ± 0.7	70.7 ± 2.6	70.3 ± 0.8
	Pair	80.7 ± 2.1	82.9 ± 4.0	75.8 ± 1.8	75.5 ± 1.8	68.9 ± 2.0	68.8 ± 2.2
KDEM	Uniform	85.5 ± 0.8	56.0 ± 5.2	56.0 ± 5.2	64.3 ± 2.0	58.6 ± 3.2	58.0 ± 2.4
	Pair	80.4 ± 3.6	58.9 ± 2.5	58.9 ± 2.5	63.6 ± 2.4	55.6 ± 5.3	56.8 ± 2.4
PPEM	Uniform	85.9 ± 1.1	56.1 ± 4.3	58.3 ± 4.8	63.6 ± 2.0	60.2 ± 0.9	58.3 ± 3.1
	Pair	80.4 ± 4.2	59.3 ± 1.5	59.5 ± 2.1	60.4 ± 4.7	51.3 ± 3.6	53.2 ± 4.9
EvGNN (Ours)	Uniform	**87.8 ± 0.8**	**86.9 ± 1.8**	**83.1 ± 1.2**	**81.2 ± 0.9**	**80.9 ± 1.6**	**72.7 ± 1.1**
	Pair	**84.9 ± 0.9**	**85.9 ± 1.6**	**82.0 ± 1.1**	**81.2 ± 1.3**	**81.3 ± 1.2**	70.6 ± 0.7

Table 2
Ablation study on six benchmark datasets (mean ± std). (Multi-expert Framework: MeF; Decision Fusion learning: DFL; Consistency Learning: ConL; Pseudo-labeling: PseL).

	MeF	DFL	ConL	PseL	Noise	Coauthor CS	Amazon Photo	Amazon Computers	Pubmed	DBLP	Citeseer
M_1	✓				Uniform	86.3 ± 1.2	84.1 ± 2.2	81.0 + 2.7	78.7 + 1.7	74.7 + 1.9	68.1 + 2.1
					Pair	82.6 ± 1.5	84.1 ± 2.7	80.7 ± 1.1	79.3 ± 1.6	77.1 ± 2.0	67.5 ± 1.1
M_2	✓	✓			Uniform	87.1 ± 0.9	85.8 ± 1.6	82.6 + 1.4	79.0 + 0.7	75.0 + 2.4	69.0 + 1.6
					Pair	82.9 ± 0.8	84.9 ± 2.1	81.3 ± 1.3	80.2 ± 1.2	77.9 ± 1.6	67.9 ± 1.3
M_3	✓	✓	✓		Uniform	87.4 ± 0.6	86.6 ± 1.6	82.9 + 1.7	79.4 ± 1.3	78.3 + 1.9	70.5 + 0.9
					Pair	83.5 ± 1.2	85.3 ± 2.0	81.5 ± 0.9	80.5 ± 1.1	79.0 ± 1.1	68.8 ± 0.9
M_4	✓	✓	✓	✓	Uniform	87.8 ± 0.8	86.9 ± 1.8	83.1 ± 1.2	81.2 ± 0.9	80.9 ± 1.6	72.7 ± 1.1
					Pair	84.9 ± 0.9	85.9 ± 1.6	82.0 ± 1.1	81.2 ± 1.3	81.3 ± 1.2	70.6 ± 0.7

Transformer-based modeling with expert merging, where KDEM enhances merging through knowledge distillation and PPEM promotes it through parameter proximity among experts. Although these methods achieve competitive performance in some settings, their results vary considerably across datasets and noise types, indicating limited stability under noisy and sparse supervision.

- Overall, our proposed EvGNN demonstrates consistently strong performance across the six benchmarks and two noise settings, achieving the best results in most cases and remaining competitive in the others. The strong overall performance of EvGNN can be attributed to three aspects: (i) The introduction of uncertainty estimation on graphs improves the confidence of predictions via a multi-expert framework; (ii) The trained high-confidence model produces pseudo-labels, which provides sufficient supervision

Table 3
Analysis of evidence-based supervised learning and cross-entropy pseudo-labeling on Amazon Photo, Pubmed, and Citeseer (Evidence-based Supervised Loss: ESL; Cross-entropy Supervised Loss: CeSL; Evidence-based Pseudo-labeling Loss: EPseL; Cross-entropy Pseudo-labeling Loss: CePseL).

ESL	CeSL	EPseL	CePseL	Noise	Amazon Photo	Pubmed	Citeseer
	✓	✓		Uniform	83.7	78.6	70.7
				Pair	83.2	79.3	67.9
✓		✓		Uniform	85.9	79.9	71.1
				Pair	83.7	80.0	68.3
	✓		✓	Uniform	84.8	80.5	71.9
				Pair	84.5	80.6	69.4
✓			✓	Uniform	86.9	81.2	72.7
				Pair	85.9	81.2	70.6

Table 4
Comparison of individual expert and expert fusion on Pubmed, DBLP, and Citeseer.

Dataset	Noise	Expert 1	Expert 2	Expert 3	EvGNN
Pubmed	Uniform	79.2	80.1	80.5	81.2
	Pair	79.1	80.6	80.4	81.2
DBLP	Uniform	80.4	80.6	80.7	80.9
	Pair	80.8	81.1	80.2	81.3
Citeseer	Uniform	71.0	72.0	72.3	72.7
	Pair	69.1	70.5	69.9	70.6

information; (iii) The consistency regularization between the representation and decision levels can fully utilize the information from unlabeled nodes to alleviate label sparsity.

4.3. Ablation study

Ablation Study on Main Components. Here we conduct an ablation analysis for the key components in our EvGNN, *i.e.*, multi-expert framework (MeF, $\mathcal{L}_{evi}$), decision fusion learning (DFL, $\mathcal{L}_{evi_fus}$), consistency learning (ConL, $\mathcal{L}_{con}$), and pseudo-labeling (PseL, $\mathcal{L}_{pse}$). We evaluate the contribution of each component by sequentially adding them to the basic model. The results are presented in Table 2. The baseline model (M_1) only contains evidence-based loss from each single expert. When considering evidential decision fusion (M_2), the performance improvement demonstrates the effectiveness of multi-view fusion. Building upon this, we further incorporate consistency learning (M_3), which results in varying degrees of improvement across all datasets, particularly on the DBLP dataset, indicating the significance of leveraging unlabeled nodes. Finally, we introduce a pseudo-labeling strategy (M_4), and further performance gain shows that pseudo-labels can provide more supervisory signals for the training process, enabling the model to learn more beneficial decision-making semantic features.

Comparison between the Evidence-based Loss and Cross-entropy Loss. To analyze the impact of evidence-based loss and traditional cross-entropy loss, we conduct an ablation study in Table 3. Comparing the first row with the second row and the third row with the fourth row shows that evidence-based supervised loss (ESL) consistently outperforms cross-entropy supervised loss (CeSL). It suggests that ESL is more suitable for supervised learning with potentially corrupted labels, since it represents predictions with Dirichlet distributions and explicitly models uncertainty, thereby reducing overconfident fitting to noisy annotations. In contrast, comparing the first row with the third row and the second row with the fourth row shows that cross-entropy pseudo-labeling loss (CePseL) performs better than evidence-based pseudo-labeling loss (EPseL). Since pseudo labels are selected from high-confidence decision-level fused evidence, CePseL can provide sharper class-discriminative supervision. Therefore, ESL and CePseL play complementary roles: the former mitigates noisy observed labels through uncertainty modeling, while the latter exploits reliable pseudo labels to provide additional supervision.

Comparison of Single Expert and Expert Fusion. To investigate the effect of fusing expert decisions, we compare the performance of the three individual experts with that of EvGNN on Pubmed, DBLP, and Citeseer under uniform and pair noise. The results in Table 4 show that the individual experts achieve different performance across datasets and noise settings, and no single expert consistently performs best. In contrast, the final fused prediction of EvGNN achieves the highest accuracy under all evaluated settings, with particularly clear gains on Pubmed. These results suggest that the experts capture complementary predictive information and the proposed fusion mechanism can effectively integrate their decisions to obtain more reliable predictions.

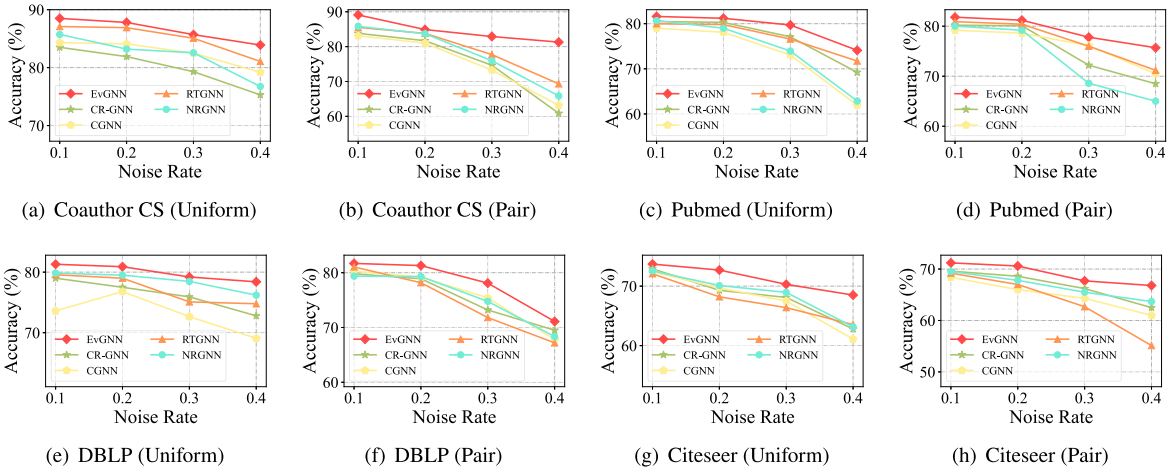


Fig. 3. Comparison *w.r.t.* different noise rates on Coauthor CS, Pubmed, DBLP, and Citeseer.

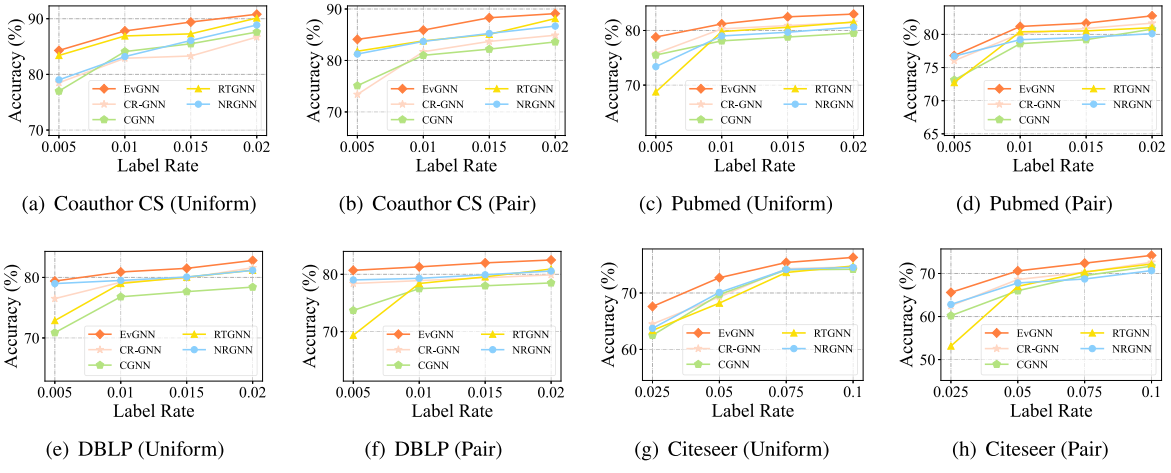


Fig. 4. Comparison *w.r.t.* different label rates on Coauthor CS, Pubmed, DBLP, and Citeseer.

4.4. Analysis on noise rate and label rate

Effect of Noise Rate. To explore the impact of different noise rates, we vary the noise rate in $\{10\%, 20\%, 30\%, 40\%\}$ under both two noise types. Here, we compare our EvGNN with four most competitive baselines (NRGNN, RTGNN, CGNN, CR-GNN) on four datasets, and the results are reported in Fig. 3. We can make the following observations:

- At different levels of noise rate, our EvGNN consistently outperforms the baselines, showing the superiority of our approach. However, the performance of the four baselines is inconsistent across different datasets, which may be due to the varying characteristics of the datasets. This also indirectly confirms that our method is able to extract the most effective semantics for decision-making on graphs.
- As the noise rate increases, the advantages of our method over the baselines become more apparent. The reason may be that both baseline methods connect labeled and unlabeled nodes, and when the noise rate is too large, most labels on the graph are incorrect. This causes erroneous signals to gradually spread throughout the graph during message propagation, resulting in significant performance degradation. However, our EvGNN fully utilizes the information of unlabeled nodes via consistency learning and alleviates the negative impact of noisy labels.

Effect of Label Rate. Here we investigate the effect of different label rates, assuming that the rates of the two noise types are both set to 20%, and then varying the label rates in $\{0.5\%, 1\%, 1.5\%, 2\%\}$ for Coauthor CS, Pubmed and DBLP datasets, and $\{2.5\%, 5\%, 7.5\%, 10\%\}$ for Citeseer dataset. From the results in Fig. 4, we can observe that:

- As the rate increases, due to sufficient supervisory signals, all methods bring significant performance gains. Among them, our EvGNN achieves optimal results on all rates, validating its superiority. Especially when the label rate is very small, EvGNN is

Table 5
Comparison on heterophilous graphs under uniform and pair noise. The best result in each column is highlighted in **bold**.

Method	Roman-Empire		Amazon-Ratings	
	Uniform	Pair	Uniform	Pair
RTGNN	25.11	25.13	35.43	33.24
RNCGLN	41.38	40.56	31.98	33.36
MEGC	25.62	25.66	33.88	30.94
KDEM	12.38	12.81	30.28	31.09
PPEM	10.26	13.60	31.26	31.33
EvGNN (Ours)	**51.87**	**51.74**	**35.74**	**35.61**

significantly higher than the four baselines, likely due to our measurement of uncertainty and sufficient utilization of unlabeled nodes.

- When comparing four baselines, it can be observed that RTGNN and CGNN perform poorly when the label rate is extremely scarce on most datasets. Regarding RTGNN, this may be due to its lower predictive accuracy when there are fewer labels, making it difficult to distinguish clean from noisy labels. This leads to an accumulation of errors during subsequent training. As for CGNN, it requires neighbor voting, which is highly inaccurate when the labels are scarce, thus resulting in significant biases. However, as the label rate increases, high-quality predictions enable the model to better overcome the influence of noisy labels and achieve better results.

4.5. Performance comparison on heterophilous graphs

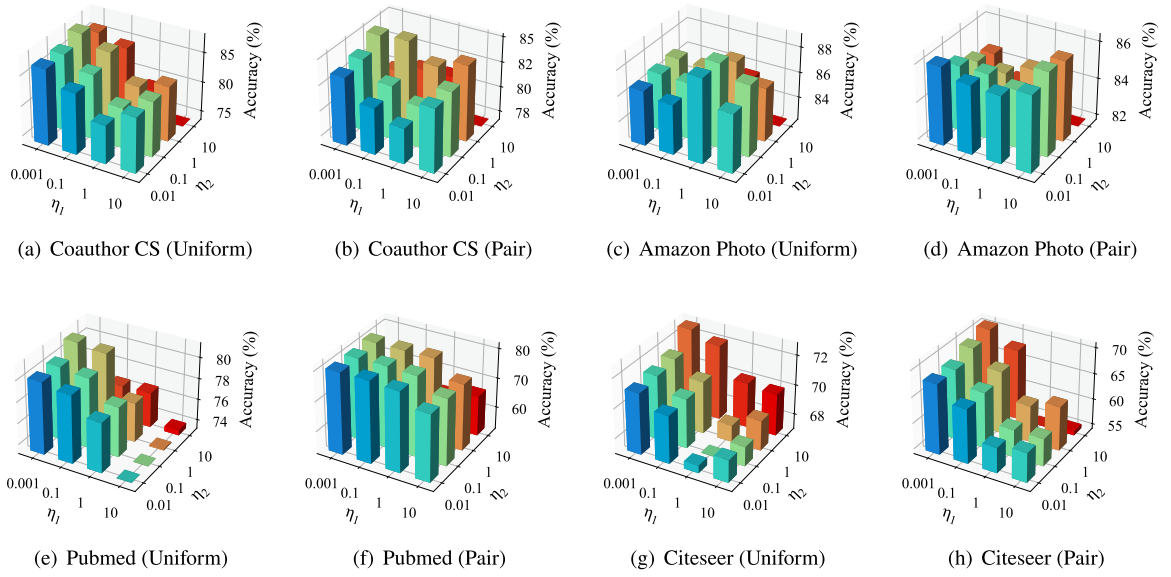
Following [74], we further evaluate the methods on Roman-Empire and Amazon-Ratings, two heterophilous graph benchmarks in which connected nodes often belong to different classes. Roman-Empire is a sparse word-dependency graph, where neighboring words frequently have different syntactic roles; Amazon-Ratings is a product co-purchasing graph, in which products that are frequently bought together may receive different discretized ratings. The results are shown in Table 5. Under 20% uniform and pair label noise with a 1% label rate, EvGNN achieves the best accuracy in all four settings, with particularly large gains on Roman-Empire.

RTGNN augments the graph and uses a consistency regularizer based on local homogeneity, which can be less reliable when connected nodes have different labels. MEGC relies on graph refinement and label propagation, whose benefits may be limited when heterophilous connections provide less informative label signals. KDEM distills an expert ensemble into a merged expert, whereas PPEM improves mergeability by bringing expert parameters closer. Their objectives focus on efficient expert merging rather than explicitly handling noisy labels or heterophilous message propagation. In EvGNN, we employ a decoupled GCN encoder that separates graph propagation from feature aggregation for heterophilous graphs, reducing harmful feature mixing between dissimilar neighbors. Together with uncertainty-aware expert fusion and the proposed regularization terms, this design enables EvGNN to maintain robust performance under noisy and sparse supervision.

4.6. Hyper-parameter analysis and complexity comparison

Impact of Loss Coefficient. We investigate the effect of the loss coefficient η_1 and η_2 in Eq. (12), which controls the ratio of the consistency learning and pseudo-labeling, respectively. We separately vary η_1 and η_2 as $\{0.001, 0.1, 1, 10\}$ and $\{0.01, 0.1, 1, 10\}$. The results of our EvGNN under two types of noise on four datasets are shown in Fig. 5. It can be observed that, generally, with the increase of η_1 while η_2 is fixed, the performance tends to decrease slowly. One possible reason is that too many unsupervised signals may overwhelm the effect of supervisory signals, resulting in erroneous biases in the gradient optimization direction. However, when η_2 is gradually increased while keeping η_1 constant, the trend is exactly the opposite. That is, with the increase of η_2, the performance tends to gradually increase. This suggests that pseudo-labeling plays a more important role in the training process, which is capable of providing sufficient supervisory signals. Nevertheless, an excessively large η_2 often leads to performance degradation, as the optimization of our pseudo-labeling is dominated by a deterministic cross-entropy loss, and a large cross-entropy loss weakens the evidence-based loss that considers uncertainty, resulting in the inability to overcome the influence of noisy labels. Therefore, it is crucial to select appropriate loss coefficients for different datasets.

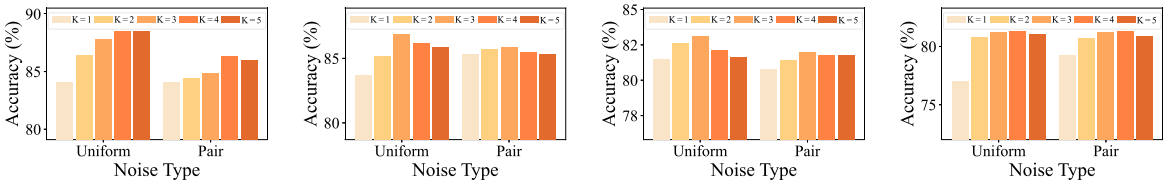
Impact of Expert Number. To explore the influence of the number of experts, we visualize the performance of four datasets on two types of noise in Fig. 6, where we vary the range of experts from 1 to 5. From the figure, we can see that as the expert number K increases from 1 to 3, there is a significant improvement in performance on all datasets. This observation suggests that introducing multiple experts effectively increases prediction diversity, enabling the fusion process to reduce predictive uncertainty and improve decision confidence through the aggregation of expert opinions. However, as K further increases, performance gradually saturates and even decreases on the Amazon Photo and Amazon Computer. A possible explanation is that, beyond a certain point, the redundant or highly similar opinions may dominate the fusion process. As a result, the fusion process gains little new information, leading to performance saturation and slight fluctuations when more experts are introduced. Moreover, too many experts also incur



(a) Coauthor CS (Uniform) (b) Coauthor CS (Pair) (c) Amazon Photo (Uniform) (d) Amazon Photo (Pair)

(e) Pubmed (Uniform) (f) Pubmed (Pair) (g) Citeseer (Uniform) (h) Citeseer (Pair)

Fig. 5. Sensitivity analysis *w.r.t.* the hyper-parameters η_1 and η_2 on Coauthor CS, Amazon Photo, Pubmed, and Citeseer.



(a) Coauthor CS (b) Amazon Photo (c) Amazon Computer (d) Pubmed

Fig. 6. Sensitivity analysis *w.r.t.* expert number K on Coauthor CS, Amazon Photo, Computer, and Pubmed.

Table 6
Comparison of computational costs on eight datasets. Each entry reports Time/epoch (s) / peak GPU Memory (MiB).

Method	Coauthor CS	Amazon Photo	Amazon Computers	Pubmed	DBLP	Citeseer	Roman-Empire	Amazon-Ratings
RTGNN	1.44/7952	0.60/3160	1.16/6888	1.52/7560	1.39/6936	0.21/904	1.74/9170	2.07/9636
RNCGLN	0.42/26712	0.09/4840	0.29/13922	0.50/25328	0.41/22940	0.02/935	0.74/33426	0.82/36734
MEGC	0.85/19480	0.45/4622	0.64/10722	0.60/13280	0.61/10566	0.27/2110	0.48/33512	0.36/23578
KDEM	1.87/1681	2.30/2223	6.12/4339	1.07/1347	1.19/1347	0.17/1347	1.52/1608	2.05/2148
PPEM	2.54/1681	2.25/2223	4.91/4250	1.31/1274	1.45/1274	0.18/1274	1.45/1462	2.30/2146
EvGNN (Ours)	0.07/7391	0.08/2412	0.09/4575	0.06/4921	0.14/9363	0.06/558	0.08/8765	0.08/10202

a computational burden, which is impractical for many critical applications. Therefore, to balance both factors, we set the number of experts to 3 to ensure both accuracy and computational efficiency.

Comparison of Computational Costs. To provide a comparison of computational efficiency, we report the average training time per epoch and peak GPU memory consumption on eight datasets. As shown in Table 6, EvGNN achieves the shortest training time per epoch on seven of the eight datasets and remains computationally efficient on Citeseer. Although its memory consumption is higher than that of KDEM and PPEM, it remains moderate and is consistently much lower than that of RNCGLN and MEGC. KDEM and PPEM have relatively low memory consumption because their memory usage is mainly determined by the fixed training batch and the SVD-based projection procedure. On Pubmed, DBLP, and Citeseer, the fixed batch is the dominant factor, resulting in nearly identical memory consumption for the two methods. However, their low memory usage comes with substantially longer training time, and does not yield consistently competitive classification performance. Overall, EvGNN offers a substantially lower time cost and a practical memory footprint while maintaining competitive classification performance.

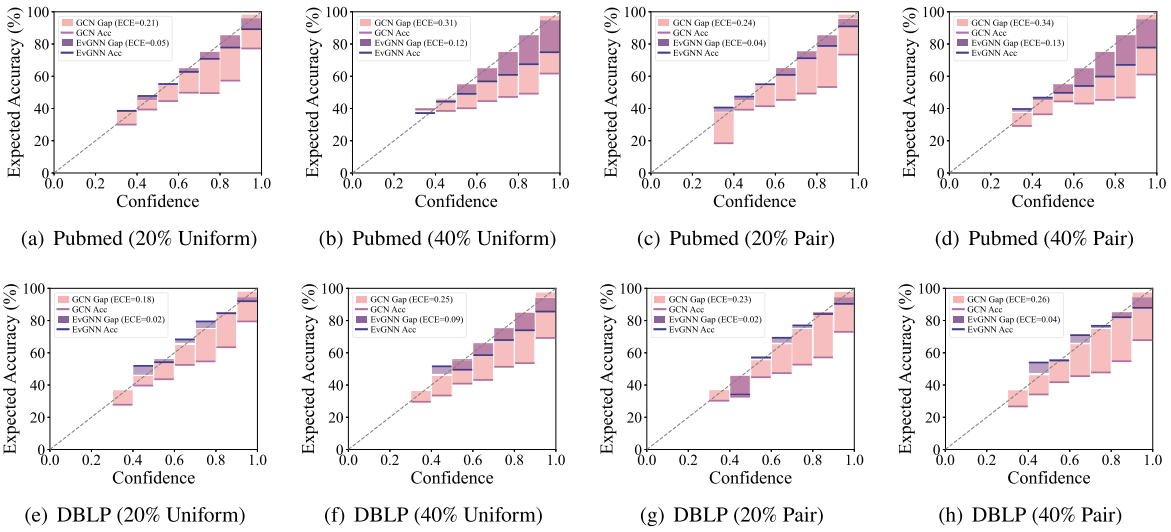


Fig. 7. Comparison of calibration gaps and ECE on Pubmed and DBLP.

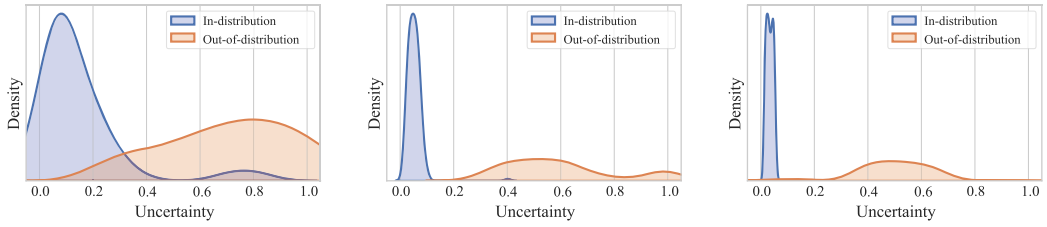


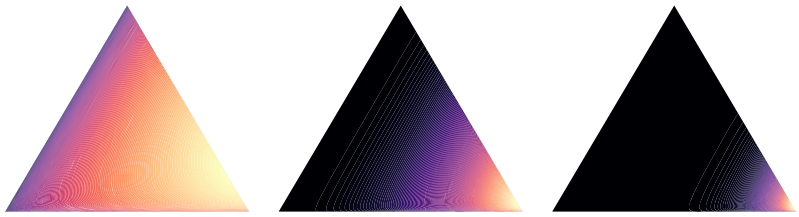
Fig. 8. Distributions of uncertainty for in-distribution and out-of-distribution data on Amazon Photo, Pubmed, and DBLP.

4.7. Analysis on calibration error and uncertainty quantification

Calibration Analysis of Test Set. Since EvGNN explicitly models uncertainty to alleviate overconfidence and overfitting, we examine how well its prediction confidence match empirical accuracy. We evaluate confidence calibration on the test set of Pubmed and DBLP using expected calibration error (ECE). The motivating experiment in the Introduction shows that the calibration of a vanilla GCN deteriorates as the noise rate increases or the label rate decreases. To examine whether EvGNN addresses this problem, Fig. 7 compares EvGNN's reliability diagrams with those of GCN. Under both uniform and pair noise, GCN exhibits pronounced gaps between confidence and empirical accuracy, particularly in the medium- and high-confidence regions. EvGNN substantially narrows these gaps and achieves a lower ECE in each setting, although the ECE of both models rises at higher noise rates. The closer agreement between confidence and empirical accuracy shows that the proposed evidence-based uncertainty learning alleviates overconfidence under noisy and sparse supervision.

Uncertainty Estimation of Out-of-distribution Data. To evaluate the effectiveness of the model in quantifying uncertainty, we conduct a comprehensive analysis of the distributions of uncertainty for both in-distribution and out-of-distribution data on three datasets. We consider the training nodes as in-distribution data, whereas we treat the test nodes with added Gaussian noise, produced by applying a fixed level of standard deviation ($\sigma = 10$), as out-of-distribution data. This approach enables us to examine the model's ability to accurately distinguish between in-distribution and out-of-distribution data. As shown in Fig. 8, the results indicate a substantial increase in uncertainty for the out-of-distribution nodes. Specifically, we observe that in the case of the Amazon Photo dataset, the estimated uncertainty for in-distribution nodes averages around 0.15, while the average uncertainty for out-of-distribution nodes exceeds 0.6. These findings suggest that our model is capable of accurately quantifying uncertainty, which can significantly facilitate the discrimination between in-distribution and out-of-distribution data.

Case Study. To illustrate the Dirichlet distribution in a triple classification task, we present a representative example using the Pubmed dataset in Fig. 9. The visualization shows the Dirichlet distribution's contours of a test sample with varying numbers of experts, providing insight into overall uncertainty and evidence over classes. We observe that when a single expert is used, it may provide weak evidence for a sample that is difficult to distinguish, resulting in relatively high uncertainty and a flat distribution over the simplex. However, as more experts are involved, the density of the distribution gradually shifts towards a specific direction and produces a sharp distribution where the density is concentrated at a vertex of a standard 2-simplex. This suggests that properly



(a) 1 expert (b) 2 experts (c) 3 experts

Fig. 9. Case study of Dirichlet distribution.

increasing the number of experts can improve the model's capability to make confident decisions, particularly when the evidence from a single expert is insufficient for uncertain samples.

5. Conclusion and discussion

In this paper, we explore an under-explored yet challenging scenario, which aims to train a robust GNN against noisy and sparse labels. Due to the label noise and costly annotation in the real world, graph labels are inevitably incorrect and scarce, leading to a poor generalization of the trained GNN. We hence propose a novel approach that incorporates uncertainty estimation into the multi-expert framework guided by evidence theory. First, to overcome the negative impact of noisy labels, we quantify the uncertainty and provide evidence for each expert by subjective logic, then we develop a fusion strategy to combine the opinions of multiple experts effectively. Moreover, to alleviate the scarcity of class labels, we introduce consistency learning and pseudo-labeling as supplemental supervision to fully leverage the unlabeled nodes. Experimental results on various benchmark datasets validate the superiority of the proposed method EvGNN.

However, our method still exhibits several limitations. For the multi-expert framework, the number of experts is empirically determined without an adaptive selection mechanism, and the current fusion strategy does not explicitly explore more flexible interaction modeling among expert representations. For pseudo-labeling, although a high-confidence threshold is used, high-confidence misclassifications remain an issue. Therefore, future work will focus on improving the adaptability and robustness of uncertainty-based GNNs, including adaptive multi-expert design with more expressive fusion and reliable pseudo-labeling strategies. Extending the framework to error detection, out-of-distribution detection, and multi-source uncertainty scenarios is also promising for safety-critical applications.

CRediT authorship contribution statement

Siyu Yi: Writing – review & editing, Writing – original draft, Software, Methodology, Funding acquisition, Conceptualization; **Wei Zhang:** Software, Validation, Visualization; **Zhengyang Mao:** Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation; **Yongdao Zhou:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization; **Ziyue Qiao:** Writing – review & editing, Software, Investigation; **Li Shen:** Writing – review & editing, Visualization, Validation, Methodology; **Dacheng Tao:** Writing – review & editing, Supervision, Conceptualization; **Jiancheng Lv:** Writing – review & editing, Supervision, Project administration, Conceptualization; **Wei Ju:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This paper is partially supported by the National Natural Science Foundation of China under Grant Numbers 12501344, 62306014, 12671309 and 12131001, the Postdoctoral Fellowship Program (Grade A) of CPSF under Grant No. BX20250376, the Sichuan Science and Technology Program under Grant Numbers 2025ZNSFSC1506 and 2025ZNSFSC0808.

Data availability

Data will be made available on request.

References

[1] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, 2017.
[2] K. Yao, J. Liang, J. Liang, M. Li, F. Cao, Multi-view graph convolutional networks with attention mechanism, Artif. Intell. 307 (2022) 103708.
[3] W. Pei, W. Xu, Z. Wu, W. Li, J. Wang, G. Lu, X. Wang, Saliency-aware regularized graph neural network, Artif. Intell. 328 (2024) 104078.
[4] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural message passing for quantum chemistry, in: International Conference on Machine Learning, PMLR, 2017, pp. 1263–1272.
[5] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Commun. ACM 64 (3) (2021) 107–115.
[6] W. Ju, Z. Liu, Y. Qin, B. Feng, C. Wang, Z. Guo, X. Luo, M. Zhang, Few-shot molecular property prediction via hierarchically structured learning on relation graphs, Neural Netw. 163 (2023) 122–131.
[7] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, M. Sugiyama, Co-teaching: robust training of deep neural networks with extremely noisy labels, Adv. Neural Inf. Process. Syst. 31 (2018) 8536–8546.
[8] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, M. Sugiyama, How does disagreement help generalization against label corruption?, in: International Conference on Machine Learning, PMLR, 2019, pp. 7164–7173.
[9] E. Malach, S. Shalev-Shwartz, Decoupling" when to update" from" how to update", Adv. Neural Inf. Process. Syst. 30 (2017) 960–970.
[10] Z. Zhang, M. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, Adv. Neural Inf. Process. Syst. 31 (2018) 8792–8802.
[11] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, L. Qu, Making deep neural networks robust to label noise: a loss correction approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1944–1952.
[12] J. Goldberger, E. Ben-Reuven, Training deep neural-networks using a noise adaptation layer, in: International Conference on Learning Representations, 2017.
[13] V.S. Sheng, J. Zhang, B. Gu, X. Wu, Majority voting and pairing with multiple noisy labeling, IEEE Trans. Knowl. Data Eng. 31 (7) (2017) 1355–1368.
[14] M. Hu, H. Han, S. Shan, X. Chen, Weakly supervised image classification through noise regularization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11517–11525.
[15] H. Song, M. Kim, J.-G. Lee, Selfie: refurbishing unclean samples for robust deep learning, in: International Conference on Machine Learning, PMLR, 2019, pp. 5907–5915.
[16] E. Dai, C. Aggarwal, S. Wang, NRGNN: learning a label noise resistant graph neural network on sparsely and noisily labeled graphs, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 227–236.
[17] S. Qian, H. Ying, R. Hu, J. Zhou, J. Chen, D.Z. Chen, J. Wu, Robust training of graph neural networks via noise governance, in: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, 2023, pp. 607–615.
[18] J. Yuan, X. Luo, Y. Qin, Y. Zhao, W. Ju, M. Zhang, Learning on graphs under label noise, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.
[19] X. Li, Q. Li, H. Qian, J. Wang, et al., Contrastive learning of graphs under label noise, Neural Netw. 172 (2024) 106113.
[20] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1321–1330.
[21] M. Hein, M. Andriushchenko, J. Bitterwolf, Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 41–50.
[22] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, J. Snoek, Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift, Adv. Neural Inf. Process. Syst. 32 (2019) 13969–13980.
[23] B. Malmir, M. Amini, S.I. Chang, A medical decision support system for disease diagnosis under uncertainty, Expert Syst. Appl. 88 (2017) 95–108.
[24] H. Zhang, X. Yang, L. Bai, J. Liang, Enhancing drug recommendations via heterogeneous graph representation learning in EHR networks, IEEE Trans. Knowl. Data Eng. 36 (7) (2023) 3024–3035.
[25] S. Liu, J. Niles-Weed, N. Razavian, C. Fernandez-Granda, Early-learning regularization prevents memorization of noisy labels, Adv. Neural Inf. Process. Syst. 33 (2020) 20331–20342.
[26] R.R. Yager, L. Liu, Classic Works of the Dempster-Shafer Theory of Belief Functions, Vol. 219, Springer, 2008.
[27] X. Luo, W. Ju, Y. Gu, Y. Qin, S. Yi, D. Wu, L. Liu, M. Zhang, Toward effective semi-supervised node classification with hybrid curriculum pseudo-labeling, ACM Trans. Multimed. Comput. Commun. Appl. 20 (3) (2023) 1–19.
[28] S. Yi, W. Ju, Y. Qin, X. Luo, L. Liu, Y. Zhou, M. Zhang, Redundancy-free self-supervised relational learning for graph clustering, IEEE Trans. Neural Netw. Learn. Syst. 35 (12) (2023) 18313–18327.
[29] W. Ju, Y. Gu, B. Chen, G. Sun, Y. Qin, X. Liu, X. Luo, M. Zhang, GLCC: a general framework for graph-level clustering, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 4391–4399.
[30] W. Ju, S. Yi, K. Zheng, Y. Wang, Z. Qiao, L. Shen, Y. Zhou, X. Cao, J. Lv, Compactness and consistency: a conjoint framework for deep graph clustering, in: The Fourteenth International Conference on Learning Representations, 2026.
[31] X. Luo, Y. Zhao, Y. Qin, W. Ju, M. Zhang, Towards semi-supervised universal graph classification, IEEE Trans. Knowl. Data Eng. 36 (1) (2023) 416–428.
[32] S.-Y. Yi, Z. Mao, W. Ju, Y.-D. Zhou, L. Liu, X. Luo, M. Zhang, Towards long-tailed recognition for graph classification via collaborative experts, IEEE Trans. Big Data 9 (6) (2023) 1683–1696.
[33] J. Bruna, W. Zaremba, A. Szlam, Y. LeCun, Spectral networks and locally connected networks on graphs, in: International Conference on Learning Representations, 2014.
[34] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 1025–1035.
[35] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, in: International Conference on Learning Representations, 2018.
[36] K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks?, in: International Conference on Learning Representations, 2019.
[37] S. Yun, M. Jeong, R. Kim, J. Kang, H.J. Kim, Graph transformer networks, in: Advances in Neural Information Processing Systems, Vol. 32, 2019, pp. 11960–11970.
[38] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, T.-Y. Liu, Do transformers really perform badly for graph representation?, in: Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 28877–28888.
[39] W. Ju, S. Yi, Y. Wang, Z. Xiao, Z. Mao, H. Li, Y. Gu, Y. Qin, N. Yin, S. Wang, et al., A survey of graph neural networks in real world: imbalance, noise, privacy and ood challenges, IEEE Trans. Pattern Anal. Mach. Intell., 48 (2025) 3036–3055.
[40] X. Du, T. Bian, Y. Rong, B. Han, T. Liu, T. Xu, W. Huang, Y. Li, J. Huang, Noise-robust graph learning by estimating and leveraging pairwise interactions, Trans. Mach. Learn. Res. (2023).
[41] J. Xia, H. Lin, Y. Xu, C. Tan, L. Wu, S. Li, S.Z. Li, GNN cleaner: label cleaner for graph structured data, IEEE Trans. Knowl. Data Eng. 36 (2) (2023) 640–651.
[42] Y. Wu, J. Yao, X. Xia, J. Yu, R. Wang, B. Han, T. Liu, Mitigating label noise on graphs via topological sample selection, in: International Conference on Machine Learning, Vol. 235 of *Proceedings of Machine Learning Research*, PMLR, 2024, pp. 53944–53972.
[43] Y. Cheng, C. Shan, Y. Shen, X. Li, S. Luo, D. Li, Resurrecting label propagation for graphs with heterophily and label noise, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 433–444.
[44] L.-H. Chen, Y. Zhang, T. Huang, L. Su, Z. Lin, X. Xiao, X. Xia, T. Liu, ERASE: error-resilient representation learning on graphs for label noise tolerance, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 2024, pp. 270–280.
[45] Z. Wang, D. Sun, S. Zhou, H. Wang, J. Fan, L. Huang, J. Bu, NoisYGL: a comprehensive benchmark for graph neural networks under label noise, in: Advances in Neural Information Processing Systems, Vol. 37, 2024, pp. 38142–38170.
[46] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059.

[47] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, Adv. Neural Inf. Process. Syst. 30 (2017) 6402–6413.
[48] W.J. Maddox, P. Izmailov, T. Garipov, D.P. Vetrov, A.G. Wilson, A simple baseline for Bayesian uncertainty in deep learning, Adv. Neural Inf. Process. Syst. 32 (2019) 13132–13143.
[49] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, et al., A review of uncertainty quantification in deep learning: techniques, applications and challenges, Inf. Fusion 76 (2021) 243–297.
[50] M. Sensoy, L. Kaplan, M. Kandemir, Evidential deep learning to quantify classification uncertainty, Adv. Neural Inf. Process. Syst. 31 (2018) 3183–3193.
[51] D.J.C. MacKay, Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks, Netw. Comput. Neural Syst. 6 (3) (1995) 469.
[52] R.M. Neal, Bayesian Learning for Neural Networks, vol. 118, Springer Science & Business Media, 2012.
[53] C. Blundell, J. Cornebise, K. Kavukcuoglu, D. Wierstra, Weight uncertainty in neural network, in: International Conference on Machine Learning, PMLR, 2015, pp. 1613–1622.
[54] C.K.I. Williams, C.E. Rasmussen, Gaussian Processes for Machine Learning, vol. 2, MIT press Cambridge, MA, 2006.
[55] R.M. Neal, Priors for Infinite Networks, in: Bayesian Learning for Neural Networks, Springer, 1996, pp. 29–53.
[56] A.P. Dempster, A generalization of Bayesian inference, J. R. Stat. Soc. B 30 (2) (1968) 205–232.
[57] A.P. Dempster, Upper and lower probabilities induced by a multivalued mapping, in: The Annals of Mathematical Statistics, 1967, p. 325–339.
[58] A. Jøsang, Subjective Logic, vol. 4, Springer, 2016.
[59] B. Škorić, S.J.A. de Hoogh, N. Zannone, Flow-based reputation with uncertainty: evidence-based subjective logic, Int. J. Inf. Secur. 15 (2016) 381–402.
[60] Z. Han, C. Zhang, H. Fu, J.T. Zhou, Trusted multi-view classification, in: International Conference on Learning Representations, 2020.
[61] S. Lee, S. Purushwalkam, M. Cogswell, D. Crandall, D. Batra, Why m heads are better than one: Training a diverse ensemble of deep networks, (2015). arXiv preprint arXiv:1511.06314
[62] S. Fort, H. Hu, B. Lakshminarayanan, Deep ensembles: a loss landscape perspective, (2019). arXiv preprint arXiv:1912.02757
[63] A.P. Dempster, et al., Upper and lower probabilities induced by a multivalued mapping, Class. Works Dempster-Shafer Theory Belief Funct. 219 (2) (2008) 57–72.
[64] F. Xiao, Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy, Inf. Fusion 46 (2019) 23–32.
[65] Z. Han, C. Zhang, H. Fu, J.T. Zhou, Trusted multi-view classification with dynamic evidential fusion, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2) (2022) 2551–2566.
[66] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[67] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C.A. Raffel, E.D. Cubuk, A. Kurakin, C.-L. Li, FixMatch: simplifying semi-supervised learning with consistency and confidence, Adv. Neural Inf. Process. Syst. 33 (2020) 596–608.
[68] D.-H. Lee, Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks, in: Workshop on Challenges in Representation Learning, ICML, Vol. 3, 2013, p. 896.
[69] H. NT, C.J. Jin, T. Murata, Learning graph neural networks with noisy labels, (2019). arXiv preprint arXiv:1905.01591
[70] M. Zhang, L. Hu, C. Shi, X. Wang, Adversarial label-flipping attack and defense for graph neural networks, in: 2020 IEEE International Conference on Data Mining (ICDM), IEEE, 2020, pp. 791–800.
[71] Y. Zhu, L. Feng, Z. Deng, Y. Chen, R. Amor, M. Witbrock, Robust node classification on graph data with graph and label noise, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 17220–17227.
[72] S. Fu, Q. Peng, Y.-M. Cheung, Y. Xu, B. Zou, X.-Y. Jing, X. You, Multiplex experts governance collaboration for label noise-resistant graph representation learning, IEEE Trans. Syst. Man Cybern. Syst. 55 (10) (2025) 7437–7448. https://doi.org/10.1109/TSMC.2025.3595183
[73] L. Liu, X. Xia, Q. Xie, B. Liu, W. Xu, M. Peng, Enhanced expert merging for mixture-of-experts in graph foundation models, in: Advances in Neural Information Processing Systems, Vol. 38, 2025, pp. 22173–22207.
[74] O. Platonov, D. Kuznedelev, M. Diskin, A. Babenko, L. Prokhorenkova, A critical look at the evaluation of GNNs under heterophily: are we really making progress?, in: International Conference on Learning Representations, 2023.