ADVERSARIALLY ROBUST GRAPH NEURAL NETWORKS VIA ADAPTIVE MULTI-FREQUENCY FILTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph Neural Networks (GNNs) have achieved remarkable success in various graph learning tasks. However, their vulnerability to adversarial attacks limits their trustworthy deployment in real-world applications. In recent years, spectral-based defense methods have provided a new perspective for enhancing the adversarial robustness of GNNs. However, existing methods typically adopt a single, fixed frequency response, which makes them difficult to adapt to spectral variations induced by adversarial attacks. Moreover, adversarial perturbations are often localized, whereas existing spectral filtering methods mainly rely on global spectral responses and lack the capability to capture localized graph variations, which compromises their adversarial robustness. To address these issues, we propose AdaMF, a dual-perspective robustness framework that captures localized graph variations and adaptively adjusts the spectral responses. Specifically, we construct learnable low-, high-, and mid-frequency filter bases with adaptive weighting coefficients to characterize diverse spectral shifts caused by adversarial perturbations, enabling flexible global spectral adaptation. Furthermore, we introduce framelet-based multi-resolution filtering to learn representations across different graph scales, enabling the model to preserve local structural details and global semantic information. Additionally, we employ prototype contrastive learning guided by pseudo-labels to further enhance the robustness of node representations. Extensive experiments on real-world datasets demonstrate that the proposed method achieves superior robustness against various adversarial attacks.

1 INTRODUCTION

Graph data provides an effective way to model real-world scenarios, where nodes represent entities and edges describe their interactions. The effective mining and utilization of the rich information contained in graph data has long been an important research direction. In recent years, graph neural networks (GNNs) Bruna et al. (2014); Kipf & Welling (2016); Wu et al. (2020) have achieved remarkable success in graph data mining, demonstrating promising application prospects in areas such as recommender systems Guo et al. (2025), financial risk control Li et al. (2025), and smart grids Sweeten et al. (2023).

However, GNNs are vulnerable Sun et al. (2022). Even unnoticeable perturbations in graph structure or node attributes can lead to a significant drop in model performance, severely hindering the safe deployment in real-world applications. For example, in smart grids Sweeten et al. (2023), where GNNs are used for power flow approximation, imperceptible perturbations to node attributes such as power demand can degrade model accuracy, leading to biased state estimation and erroneous dispatch decisions. Such errors can disrupt the balance between power supply and demand, and even trigger large-scale blackouts. This vulnerability has raised serious concerns about the security of GNNs, highlighting the urgent need to develop robust methods resistant to adversarial perturbations.

Therefore, researchers have proposed various defense strategies, among which spectral analysis-based methods offer a new perspective for improving adversarial robustness. Spectral analysis is an important approach in GNNs that characterizes the spectral properties of graph data and reveals signal variations that are difficult to analyze in the spatial domain. It can be observed from related

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

defense studies that: (1) existing methods often rely on predefined frequency responses, which limits their ability to capture the diverse spectral shifts across frequency bands caused by adversarial attacks; (2) spectral response designs are often customized for specific attack strategies or datasets, which severely limits their generalization capability; (3) most methods model graph signals only from a single-scale perspective and lack the ability to jointly capture multi-scale spectral information and localized graph variations.

To address the above issues, we propose AdaMF, an adversarially robust framework that jointly models spectral variations and multi-scale graph information. AdaMF enables robust representation learning by integrating adaptive spectral modeling with framelet-based multi-resolution analysis. From the spectral perspective, we formulate the graph filtering process using multiple frequency components and adaptively learn their combination weights to accommodate diverse spectral variations. From the multi-scale perspective, framelets are employed to perform multi-resolution decomposition of graph signals, enabling the capture of local details and global semantic information. Furthermore, a pseudo-label-guided prototype contrastive objective is introduced to improve representation consistency across augmented views. By jointly optimizing the InfoNCE loss and cross-entropy loss, AdaMF learns node representations that are robust to adversarial perturbations. Extensive experiments on multiple real-world datasets demonstrate that AdaMF consistently outperforms existing defense methods under various adversarial attacks. The main contributions of this paper are summarized as follows:

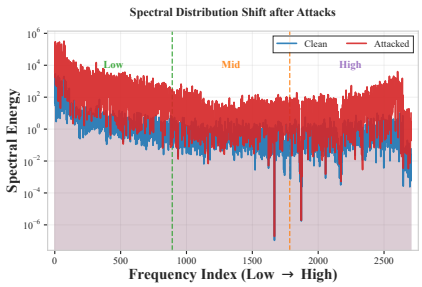


Figure 1: Spectral energy distributions of the GCN before and after adversarial attacks. The results show that adversarial attacks cause spectral distribution shifts across the entire spectrum rather than being confined to a single frequency region.

- We propose AdaMF, an adversarially robust framework that jointly captures spectral variations and multi-scale graph structures, providing complementary representations from global spectral and local scale perspectives.
- We design an adaptive multi-frequency response mechanism to capture heterogeneous spectral variations and provide theoretical insights into the perturbation sensitivity of the proposed adaptive filtering mechanism, while introducing framelet-based multi-resolution analysis into adversarial graph learning to preserve localized structural information.
- Extensive experiments on four benchmark datasets, seven adversarial attacks, and ten state-of-the-art baseline methods demonstrate the effectiveness of the proposed method, which achieves superior performance under attacks of different strengths.

2 Related Work

2.1 Adversarial Attacks

GNNs are vulnerable to adversarial attacks. According to the stage at which they are conducted, existing graph adversarial attacks can be categorized into poisoning attacks and evasion attacks. **Poisoning attacks** are conducted during the training stage, aiming to perturb the training data so that the model learns incorrect representations. Nettack Zügner et al. (2018) and Metattack Zügner & Gunnemann (2019) generate adversarial perturbations using greedy search and meta-gradients, respectively. **Evasion attacks** are conducted during the testing stage, where carefully crafted adversarial examples are generated to mislead the model without modifying its parameters. RL-S2V Dai et al. (2018) is the first work that introduces reinforcement learning into graph adversarial attacks. RWCS Ma et al. (2020) and VoteAttack Wen et al. (2024) select attack nodes based on PageRank-like Page et al. (1999) importance scores and a voting matrix, respectively. In addition, several studies investigate evasion attacks from a node injection perspective. AFGSM Wang et al. (2020) and GZOO Yu et al. (2024) are representative works that study graph adversarial attacks under the node injection setting, where adversaries are only allowed to inject malicious nodes without modifying the original graph data. Evasion attacks have attracted widespread attention due to their ease of implementation, strong stealthiness, and high destructiveness, posing significant challenges to

the security of real-world GNN deployments. Consequently, this paper focuses on improving the adversarial robustness of GNNs against evasion attacks.

2.2 Adversarial Defenses

Extensive efforts have been devoted to improving the robustness of GNNs. Existing defense methods can be broadly classified into spatial-based and spectral-based defenses. **Spatial-based** defense methods primarily suppress the propagation of adversarial perturbations through graph structure enhancement, model architecture design, and message-passing mechanisms. RGCN Zhu et al. (2019) represents each node as a Gaussian distribution to model uncertainty and mitigate adversarial perturbations. GCN-Jaccard Wu et al. (2019) and GNNGuard Zhang & Zitnik (2020) identify and remove fake edges connecting dissimilar nodes based on similarity measures. Elastic Liu et al. (2021) introduces an elastic message-passing mechanism. HSIC-GAD Guo et al. (2026) enhances model robustness by exploiting inter-layer similarity via the Hilbert-Schmidt Independence Criterion. **Spectral-based** defense methods primarily mitigate the effects of adversarial perturbations through low-rank modeling and filter design. ProGNN Jin et al. (2020) and GCN-SVD Entezari et al. (2020) remove adversarial perturbations by preserving the low-rank information of the graph adjacency matrix. Mid-GCN Huang et al. (2023) observes that mid-frequency signals are insensitive to adversarial attacks and designs a mid-pass filter to learn robust representations. ASF-GCN Fan et al. (2025) adaptively adjusts the filter shape based on graph statistics, such as the degree distribution and spectral information. In contrast, AdaMF integrates adaptive multi-frequency filtering with framelet-based multi-resolution analysis to learn robust node representations by jointly exploiting frequency and scale information.

3 Preliminaries

3.1 Notations

Given an undirected graph $\mathcal{G} = (\mathbf{A}, \mathbf{X})$ with N nodes, where $\mathbf{A} \in \{0,1\}^{N\times N}$ is the symmetric adjacency matrix and $\mathbf{X} \in \mathbb{R}^{N\times d}$ is the node feature matrix, we denote the node set as $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$. Here, $\mathbf{A}_{ij} = 1$ indicates an edge between nodes v_i and v_j, while $\mathbf{A}_{ij} = 0$ indicates no edge. Each node is associated with a d-dimensional feature vector. The degree matrix $\mathbf{D}$ is defined as $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$. The normalized graph Laplacian is defined as $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$. For node classification, $\mathcal{V}$ is partitioned into disjoint labeled nodes $\mathcal{V}_L$ and unlabeled nodes $\mathcal{V}_U$, where the labels of nodes in $\mathcal{V}_L$ belong to the label space $\mathcal{Y} = \{1, 2, \ldots, K\}$ and K denotes the number of classes.

3.2 Graph Adversarial Robustness

The goal of node classification for GNNs is to learn a model $f_{\boldsymbol{\theta}}(\cdot)$ using labeled nodes $\mathcal{V}_L$ to efficiently predict the labels of unlabeled nodes $\mathcal{V}_U$ in the graph. The adversarial robustness of this model refers to its ability to maintain high prediction accuracy even when subjected to adversarial attacks. The robustness evaluation under adversarial attacks can be formulated as:

$$E_{\hat{\mathcal{G}}\sim\Phi(\mathcal{G},\mathcal{B})} \sum_{v_j\in\mathcal{V}_U} \mathcal{L}\left(f_{\boldsymbol{\theta}^*}\left(\hat{\mathcal{G}}\right)_j, \mathcal{Y}_j\right), s.t.\ \boldsymbol{\theta}^* = arg\min_{\boldsymbol{\theta}} \sum_{v_i\in\mathcal{V}_L} \mathcal{L}_{train}\left(f_{\boldsymbol{\theta}}(\mathcal{G})_i, \mathcal{Y}_i\right), \quad (1)$$

where $\boldsymbol{\theta}$ denotes the model parameters, $\Phi(\cdot,\cdot)$ denotes the perturbation strategy function, $\hat{\mathcal{G}}$ is the perturbed graph, $\mathcal{Y}_i$ and $\mathcal{Y}_j$ denote the labels of the i-th and j-th nodes, respectively. To make the perturbations more imperceptible, attackers are usually constrained by a perturbation budget $\mathcal{B}$.

3.3 Graph Framelets

Graph framelets Zheng et al. (2021) provide a multiresolution analysis framework for graph signals. They consist of a filter bank $\eta = \left\{a; b^{(1)}, \cdots, b^{(n)}\right\}$ and their corresponding scaling functions $\Psi = \left\{\alpha; \beta^{(1)}, \cdots, \beta^{(n)}\right\}$, where a and $b^{(i)}$ denote the low-pass and the i-th high-pass filters, respectively. The low-pass filter captures smooth approximation components, while the high-pass filters

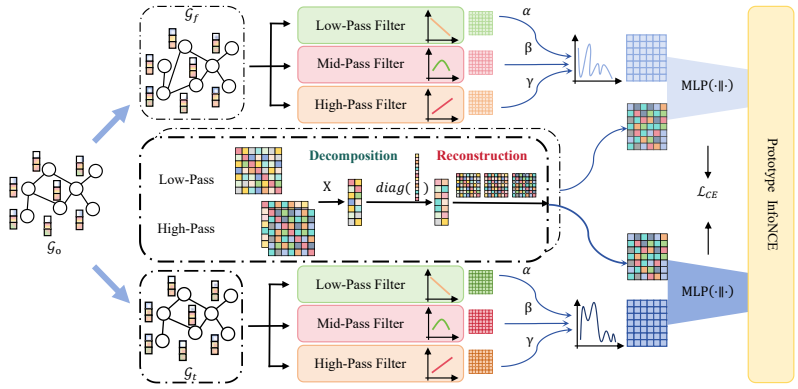


Figure 2: The overall framework of AdaMF.

extract detailed variations of graph signals. For $\gamma \in \mathbb{R}$, these filters satisfy the classical refinement conditions in the Fourier domain:

$$\hat{\alpha}(2\gamma) = \hat{a}(\gamma)\,\hat{\alpha}(\gamma), \qquad \hat{\beta}^{(i)}(2\gamma) = \hat{b}^{(i)}(\gamma)\,\hat{\alpha}(\gamma), i = 1, \cdots, n. \tag{2}$$

Suppose $\{(u_p, \lambda_p)\}_{p=1}^{N}$ are the eigenvector and eigenvalue pairs for graph Laplacian $\mathbf{L}$ of graph $\mathcal{G}$ with N nodes. At dilation level $j = 1, \cdots, J$ and for a node $q \in \mathcal{V}$, the framelets are defined as:

$$\varphi_{j,q}(v) = \sum_{p=1}^{N} u_p(v)\,\hat{\alpha}\left(\frac{\lambda_p}{2^j}\right) u_p(q), \qquad \psi_{j,q,i}(v) = \sum_{p=1}^{N} u_p(v)\,\hat{\beta}^{(i)}\left(\frac{\lambda_p}{2^j}\right) u_p(q), \tag{3}$$

where $\varphi_{j,q}$ and $\psi_{j,q,i}$ are the low-pass and i-th high-pass framelets centered at node q, respectively. The commonly used Haar-type filter bank Dong (2017) can ensure tightness and perfect reconstruction. Given a graph signal $f \in \mathbb{R}^N$, the framelet coefficients Li et al. (2026) are computed as:

$$\mathcal{W}_J^0 f = \mathbf{U}\hat{a}\left(\frac{\Lambda}{2}\right)\mathbf{U}^T f, \qquad \mathcal{W}_j^i f = \mathbf{U}\hat{b}^{(i)}\left(\frac{\Lambda}{2^{j+1}}\right)\mathbf{U}^T f, \tag{4}$$

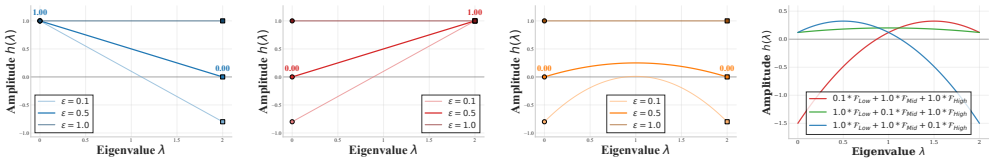
where $\mathcal{W}_J^0 f$ and $\mathcal{W}_j^i f$ are the low-pass and high-pass coefficients of f. Since eigen-decomposition is computationally expensive for large graphs, framelet operators can be approximated by Chebyshev polynomial expansions: $\hat{\mathcal{W}}_j^1 = \mathcal{T}_r\left(2^{-M}\mathbf{L}\right)$, $\hat{\mathcal{W}}_j^i = \mathcal{T}_r\left(2^{-M-j+1}\mathbf{L}\right)\mathcal{T}_0\left(2^{-M-j+2}\mathbf{L}\right)\cdots\mathcal{T}_0\left(2^{-M}\mathbf{L}\right)$, where $\mathcal{T}_r$ is the r-th order Chebyshev polynomial. To ensure that the scaled spectrum falls within the valid frequency range, a constant M is selected such that $\lambda_{max} \leq 2^M\pi$, where λ_{max} is the largest eigenvalue of $\mathbf{L}$.

4 Proposed Method: AdaMF

The overall architecture of the proposed **AdaMF** framework is illustrated in Figure 2. AdaMF is an adversarially robust graph learning framework that integrates adaptive multi-frequency filtering with framelet-based multi-resolution filtering to capture complementary spectral and multi-scale information.

4.1 Automatic Frequency Filtering

We analyze the spectral variations of GCN representations under the representative attack VoteAttack Wen et al. (2024). As shown in Figure 1, experiments on Cora show that VoteAttack perturbs the entire frequency spectrum rather than a specific frequency band. Consequently, **a single, fixed frequency response is insufficient to adapt to the diverse spectral variations induced by adversarial perturbations**. To overcome this limitation, we propose a hybrid filter consisting of low-pass, high-pass, and mid-pass components, whose adaptive combination enhances robustness. We first formulate each base filter and then describe their joint design:



(a) Low-pass Filter (b) High-pass Filter (c) Mid-pass Filter (d) Hybrid Filter

Figure 3: The relations between eigenvalues and amplitudes in different filters.

- The low-pass filter is $\mathcal{F}_L = \varepsilon\mathbf{I} + (1-\varepsilon)\,\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - (1-\varepsilon)\,\mathbf{L}$. As shown in Figure 3(a), the spectral response of the low-pass filter decreases monotonically with increasing eigenvalue λ, indicating that higher-frequency components receive progressively smaller weights. Consequently, the filter effectively preserves low-frequency information while suppressing high-frequency components, thereby smoothing graph signals to capture the shared characteristics of node features.
- The high-pass filter is $\mathcal{F}_H = \varepsilon\mathbf{I} - (1-\varepsilon)\,\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\,\mathbf{D}^{-\frac{1}{2}} = (2\varepsilon-1)\,\mathbf{I} + (1-\varepsilon)\,\mathbf{L}$. As shown in Figure 3(b), the high-pass filter exhibits an increasing spectral response with respect to λ, opposite to the decreasing trend of the low-pass filter. High-frequency signals correspond to detailed information in graphs and are commonly used to capture differences in graph signals.
- The mid-pass filter is $\mathcal{F}_M = (2\varepsilon-1)\,\mathbf{I} + 2\,(1-\varepsilon)^2\,\mathbf{L} - (1-\varepsilon)^2\,\mathbf{L}^2$. As shown in Figure 3(c), the spectral response of the mid-pass filter first increases and then decreases with respect to the Laplacian eigenvalue λ, reaching its maximum at $\lambda = 1$. The filter enhances middle-frequency components while suppressing both low-frequency and high-frequency components, thereby selectively preserving information within the target frequency band.
- The hybrid filter is $\mathcal{F}_{Hybrid} = \alpha\mathcal{F}_L + \beta\mathcal{F}_H + \gamma\mathcal{F}_M$, where α, β, and γ are learnable parameters constrained to satisfy $\alpha+\beta+\gamma = 1$ and $\alpha, \beta, \gamma \geq 0$. These coefficients are jointly optimized with the model parameters, enabling the model to adaptively capture informative frequency components from graph signals. As shown in Figure 3(d), different parameter combinations produce distinct spectral responses. Theorem 1 demonstrates that the hybrid filter of AdaMF exhibits controllable sensitivity to perturbations, while Theorem 2 further establishes the perturbation stability of the adaptive spectral filtering operator, providing a theoretical explanation for its robustness. Furthermore, perturbation-based augmentations during training encourage the spectral coefficients to learn more stable frequency responses, allowing the model to effectively balance different frequency components and improve adversarial robustness.

The hybrid filter is shared across all nodes in the graph, and the node representations are defined as follows:

$$\mathbf{Z}_1^{(l+1)} = \sigma\left(\mathcal{F}_{Hybrid}\left(\hat{\mathbf{A}}\right)\mathbf{Z}_1^{(l)}\mathbf{W}_1^{(l)}\right), \tag{5}$$

where $\sigma(\cdot)$ denotes an activation function, $\hat{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, $\mathbf{Z}_1^{(0)} = \mathbf{X}$, $\mathbf{W}_1^{(l)}$ denotes the learnable parameters of the l-th layer.

We now characterize the spectral family actually searched by AdaMF and analyze its sensitivity to attack-induced graph perturbations. Let $q = 1-\varepsilon$, $\mathbf{P} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, and $\mathbf{L} = \mathbf{I} - \mathbf{P}$, where $\mathbf{L}$ has eigenvalues in $[0, 2]$. Since $\mathcal{F}_{\text{Hybrid}}$ is a polynomial in L of degree at most two (equivalently, in P), and $P = \hat{A}$ as defined in Section 4.1, we write $\mathcal{F}_{\text{Hybrid}}(L)$ in what follows, interchangeably with $\mathcal{F}_{\text{Hybrid}}(\hat{A})$. Recall that the learnable coefficients satisfy $(\alpha, \beta, \gamma) \in \Delta_2$, where $\Delta_2 = \{(\alpha, \beta, \gamma) \in \mathbb{R}_+^3 : \alpha + \beta + \gamma = 1\}$ is the probability simplex. The spectral responses of the three base filters are

$$g_L(\lambda) = 1 - q\lambda, \qquad g_H(\lambda) = 2\varepsilon - 1 + q\lambda, \qquad g_M(\lambda) = \varepsilon^2 - q^2(\lambda-1)^2, \tag{6}$$

the hybrid response family is $\mathcal{R}_\varepsilon = \{\alpha g_L + \beta g_H + \gamma g_M : (\alpha, \beta, \gamma) \in \Delta_2\}$, and the spectral sensitivity of a response r is measured by $\mathcal{J}(r) = \max_{\xi\in[0,2]} |r'(\xi)|$. A distinguished sub-family is $r_\gamma = (1-\gamma)\varepsilon + \gamma g_M$, obtained for $\alpha = \beta = (1-\gamma)/2$, which satisfies $r'_\gamma(1) = 0$; the structural properties of $\mathcal{R}_\varepsilon$ and r_γ are established in Propositions 1–2 (Appendix A).

Theorem 1 (Spectral perturbation bounds)**.** *Let $\lambda, \widetilde{\lambda} \in [0,2]$ satisfy $|\widetilde{\lambda} - \lambda| \leq \delta$. Then, for every $r \in \mathcal{R}_\varepsilon$, $|r(\widetilde{\lambda}) - r(\lambda)| \leq \mathcal{J}(r)\,\delta$. Moreover, for every $\gamma \in [0,1]$, $r_\gamma(\widetilde{\lambda}) - r_\gamma(\lambda) = 2q^2\gamma(1-\lambda)(\widetilde{\lambda} - \lambda) - \gamma q^2(\widetilde{\lambda} - \lambda)^2$; and, in particular, for every $\widetilde{\lambda} \in [0,2]$, $r_\gamma(\widetilde{\lambda}) - r_\gamma(1) = -\gamma q^2(\widetilde{\lambda} - 1)^2$.*

For a scalar response r, we write $\mathcal{F}(\mathbf{L}) = \mathbf{U} r(\mathbf{\Lambda})\mathbf{U}^\top$ for the induced matrix filter, and abbreviate $\mathcal{F}_\gamma(\mathbf{L}) := r_\gamma(L)$.

Theorem 2 (Stability under graph perturbations)**.** *Let $\widetilde{\mathbf{L}}$ be the normalized Laplacian of the attacked graph, let $\Delta = \widetilde{\mathbf{L}} - \mathbf{L}$, and let $\eta = \|\Delta\|_2$. Then, for every $(\alpha, \beta, \gamma) \in \Delta_2$, $\left\|\mathcal{F}_{\mathrm{Hybrid}}(\widetilde{\mathbf{L}}) - \mathcal{F}_{\mathrm{Hybrid}}(\mathbf{L})\right\|_2 \leq \left[q(\alpha+\beta) + 2q^2\gamma\right]\eta$, and, for every $\gamma \in [0,1]$, $\left\|\mathcal{F}_\gamma(\widetilde{\mathbf{L}}) - \mathcal{F}_\gamma(\mathbf{L})\right\|_2 \leq 2q^2\gamma\,\eta$.*

4.2 Multi-Scale Framelet Filtering

Since adversarial attacks usually have limited perturbation budgets, attackers can only modify a small portion of nodes, edges, or features, and the resulting adversarial perturbations typically exhibit localized characteristics. Although the hybrid filter enables flexible spectral adaptation, it mainly focuses on global spectral modeling and may **fail to fully capture localized graph variations and multi-scale structural information**. To address this issue, we further introduce framelet filtering with multi-scale analysis capability, which enhances the representation of graph signals across different resolutions. Moreover, framelets Zheng et al. (2021) can reduce both feature and structure noise. It is represented as follows:

$$\mathbf{Z}_2 = \sigma\left(\mathcal{V}\left(diag\left(\boldsymbol{\theta}\right)\hat{\mathcal{W}}\mathbf{X}\mathbf{W}_2\right)\right), \tag{7}$$

where $\mathcal{V} \in \mathbb{R}^{N\times(nJ+1)N}$ is reconstruction operator, $diag\left(\boldsymbol{\theta}\right) \in \mathbb{R}^{(nJ+1)N\times(nJ+1)N}$ denotes the diagonal matrix of learnable framelet coefficients, $\hat{\mathcal{W}} \in \mathbb{R}^{(nJ+1)N\times N}$ denotes the framelet decomposition operator approximated by Chebyshev polynomials, $\mathbf{W}_2 \in \mathbb{R}^{d\times c}$ is a trainable feature transformation matrix.

4.3 Frequency-Aware Fusion

Although the hybrid filter and framelet filtering capture complementary spectral information from global and multi-scale perspectives, respectively, each representation alone may be insufficient for robust node classification under adversarial attacks. Therefore, we fuse these representations to exploit their complementary frequency characteristics. Furthermore, considering that adversarial attacks introduce perturbations into graph data, we separately apply 1% perturbations to node attributes and graph topology to construct two augmented views. Then, contrastive learning is performed between these two views to learn consistent and robust node representations against adversarial perturbations.

The representations from the hybrid filter and framelet filtering are concatenated and transformed to obtain the fused representation:

$$\mathbf{Z}_c = \mathbf{Z}_1 \| \mathbf{Z}_2, \qquad \mathbf{Z}_f = \sigma\left(\mathbf{Z}_c\mathbf{W}_f + b\right), \tag{8}$$

where $\mathbf{W}_f \in \mathbb{R}^{2c\times K}$ and $b \in \mathbb{R}^K$ are the learnable parameters, $\mathbf{Z}_1 \in \mathbb{R}^{N\times c}$ and $\mathbf{Z}_2 \in \mathbb{R}^{N\times c}$ are the representations obtained from the hybrid filter and framelet filter, respectively. For each view, we apply the cross-entropy loss $\mathcal{L}_{ce}$ to optimize the node classification objective, based on which pseudo-labels are generated.

Given two augmented graphs $\mathcal{G}_f$ and $\mathcal{G}_t$, the representations $\hat{\mathbf{Z}}_c$ and $\tilde{\mathbf{Z}}_c$ are obtained from the two views, respectively. The prototype-level InfoNCE Tan et al. (2022) loss is defined as follows:

$$\mathcal{L}_{pcl} = -\frac{1}{K}\sum_{i=1}^{K}\log\frac{exp\left(S_{ii}/\tau\right)}{\sum_{j=1}^{K}exp\left(S_{ij}/\tau\right)}, \tag{9}$$

where τ is a temperature parameter, $S_{ij} =< \hat{P}_i, \tilde{P}_j >$ denotes the cosine similarity between the prototype $\hat{P}_i$ of the i-th class in the first augmented view and the prototype $\tilde{P}_j$ of the j-th class in the second augmented view. The class prototypes are calculated as follows:

$$P_i = \frac{\sum_{j=1}^{N} c_j \mathbb{I}\left(\hat{y}_j = i\right) z_c^j}{\sum_{j=1}^{N} c_j \mathbb{I}\left(\hat{y}_j = i\right)}, \tag{10}$$

where $\mathbb{I}(\cdot)$ represents the indicator function, $\hat{y}_j$, c_j and z_c^j denote the pseudo-label, confidence score, and node representation of the j-th node, respectively. Finally, the overall optimization objective is defined as

$$\mathcal{L} = \mathcal{L}_{ce} + \xi \mathcal{L}_{pcl}, \tag{11}$$

where ξ controls the trade-off between prototype-level contrastive learning and node-level supervision, $\mathcal{L}_{ce}$ denotes the cross-entropy loss on the two augmented views.

4.4 Time Complexity Analysis

The time complexity of the proposed model is mainly determined by the filtering operations, with a complexity of $\mathcal{O}\left(H\left(Ed + Nd^2\right)\right)$, where E and H denote the number of edges in the graph and the number of layers in the model, respectively. Moreover, the prototype-level contrastive learning introduces an additional computational complexity of $\mathcal{O}\left(NKd + K^2d\right)$. Since $K \ll N$ and $d \ll N$ in most practical scenarios, the proposed method maintains a relatively low computational complexity.

5 Experiments

To verify the effectiveness of the proposed method under adversarial attacks, we conduct experiments to investigate the following research questions:

RQ1: How does AdaMF perform in terms of adversarial robustness under different attack scenarios?

RQ2: What are the contributions of each component of AdaMF to adversarial robustness?

RQ3: How sensitive is AdaMF to key hyperparameters?

5.1 Experiment Settings

5.1.1 Datasets

We evaluate the proposed method on four publicly available benchmark datasets: Cora Yang et al. (2016), Citeseer, Pubmed, and ogbn-arxiv Hu et al. (2020). More details about these datasets are presented in Appendix D.0.1.

5.1.2 Baseline and Adversarial Attack Methods

We evaluate the proposed method against ten baselines, which are grouped into three categories: (1) classical GNNs, including GCN Kipf & Welling (2016), GAT Veličković et al. (2017), and JKNet Xu et al. (2018); (2) spatial-domain defense methods, including RGCN Zhu et al. (2019), Elastic Liu et al. (2021), and HSIC-GAD Guo et al. (2026) (abbreviated as GAD); and (3) spectral-domain defense methods, including SVD Entezari et al. (2020), AutoGCN Wu et al. (2022), Mid-GCN Huang et al. (2023) (abbreviated as Mid), and ASF-GCN Fan et al. (2025). Moreover, we employ seven adversarial attack methods from two categories: (1) graph modification attacks (GMAs), including GC-RWCS Ma et al. (2020), RWCS, InfMax-Norm Ma et al. (2022), InfMax-Unif, and VoteAttack Wen et al. (2024); and (2) graph injection attacks (GIAs), including AFGSM Wang et al. (2020) and GZOO Yu et al. (2024). A detailed introduction to baselines and attack methods can be found in Appendices D.0.2 and D.0.3, respectively.

5.1.3 Implementation Details

We employ a two-layer model optimized by Adam Kingma & Ba (2014) for 200 epochs, with a learning rate of 0.01, a dropout rate of 0.5, a weight decay of 5×10^{-5}, and a hidden dimension of 16. The hyperparameter ξ is selected from the candidate set $\{0.1, 0.3, 0.5, 0.7, 0.9, 1\}$, and $\xi = 1$ is adopted in all experiments since the model is insensitive to this parameter. The dataset is split according to the attack settings, with a 60%/20%/20% train/validation/test split for GMAs and a 10%/10%/80% split for GIAs. In GMAs, 1% of nodes are selected as attack targets, and each target node is perturbed with an intensity of 1 on 2% of its feature dimensions. In GIAs, test nodes are treated as attack targets, except for AFGSM, which selects a fixed set of 50 nodes. All experiments

Table 1: Node classification accuracy (%) under adversarial attacks. The best and second-best results are highlighted in bold and underlined, respectively. OOM denotes out-of-memory.

Dataset	Method	GCN	GAT	JKNet	SVD	RGCN	Elastic	AutoGCN	Mid	ASF-GCN	GAD	Ours
Cora	Clean	83.03	85.73	82.01	80.57	85.77	84.82	85.40	82.26	86.32	87.42	**88.17**
	GC-RWCS	74.49	82.62	42.51	67.51	82.07	79.30	85.03	72.27	85.77	85.76	**87.99**
	RWCS	75.57	83.36	49.72	65.67	82.25	79.30	85.03	73.38	86.14	85.95	**88.35**
	InfMax-Unif	73.01	81.89	43.25	64.88	82.25	76.71	85.03	72.09	85.95	85.95	**87.43**
	InfMax-Norm	73.94	82.81	42.33	64.01	81.88	78.56	85.03	73.01	85.95	85.39	**87.80**
	VoteAttack	73.01	82.25	42.51	65.71	82.99	78.74	84.84	74.68	85.77	85.95	**87.99**
	AFGSM	36.00	20.00	54.00	60.00	16.00	50.00	58.00	74.00	70.00	25.00	**76.00**
	GZOO	19.53	68.74	32.78	67.75	64.23	60.68	19.21	43.60	70.23	72.03	**74.01**
Citeseer	Clean	74.89	73.08	73.21	70.00	78.80	74.44	75.19	71.28	75.42	78.50	**79.20**
	GC-RWCS	52.78	70.83	45.71	59.55	74.28	54.89	73.68	58.64	73.68	78.35	**78.47**
	RWCS	54.89	69.62	47.82	55.89	71.42	58.50	74.14	61.20	73.08	78.50	**79.10**
	InfMax-Unif	54.59	71.58	47.97	59.06	73.38	55.64	73.83	59.55	73.53	78.50	**79.10**
	InfMax-Norm	53.38	71.13	45.71	57.28	73.83	55.64	73.83	58.50	73.23	78.35	**78.50**
	VoteAttack	50.23	70.38	47.37	56.91	73.08	55.64	73.98	58.65	73.23	78.35	**78.42**
	AFGSM	40.00	40.00	50.00	64.00	28.00	52.00	44.00	64.00	60.00	20.00	**66.00**
	GZOO	35.82	52.33	24.49	32.16	33.35	22.50	20.55	33.61	57.96	70.93	**73.69**
Pubmed	Clean	87.95	85.20	86.29	84.07	85.04	86.25	**89.78**	84.39	87.01	85.65	88.56
	GC-RWCS	64.82	70.88	47.25	57.04	79.12	58.66	73.75	71.59	84.02	81.59	**87.29**
	RWCS	67.61	67.38	48.26	57.61	77.49	56.30	74.77	68.36	81.87	80.07	**86.36**
	InfMax-Unif	64.95	71.87	46.49	60.76	78.64	58.74	74.64	72.14	83.72	81.18	**87.62**
	InfMax-Norm	62.59	71.18	46.01	58.34	78.87	57.80	73.90	71.78	83.31	81.31	**87.62**
	VoteAttack	63.65	65.91	48.21	53.28	76.29	56.81	70.02	69.74	78.65	79.93	**86.43**
	AFGSM	52.00	54.00	62.00	86.00	46.00	68.00	54.00	70.00	78.00	32.50	**90.00**
	GZOO	25.86	33.20	15.15	27.32	21.32	30.82	18.39	33.68	41.37	16.14	**67.87**
ogbn-arxiv	Clean	48.86	46.27	**62.57**	OOM	33.98	25.86	49.32	26.08	60.23	33.66	62.39
	VoteAttack	46.72	43.68	49.83	OOM	26.47	17.55	44.51	23.94	45.27	28.48	**60.70**

run on a server equipped with an Intel® Xeon® Gold 6254 CPU @ 3.10GHz and NVIDIA GeForce RTX 3090 GPU.

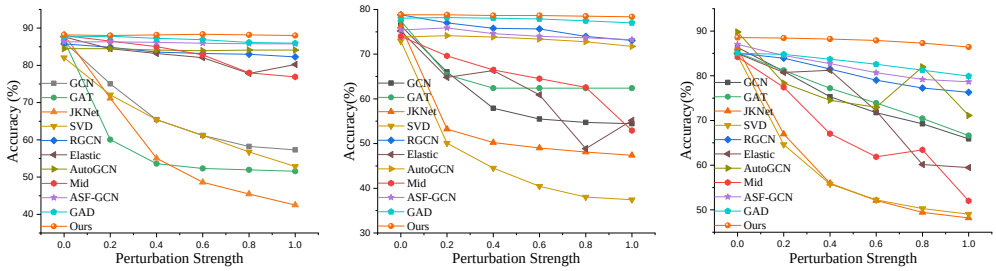
5.2 Performance under Adversarial Attacks (RQ1)

We use the prediction accuracy as the evaluation metric. As shown in Table 1, the proposed method achieves competitive clean performance and the best robustness among all baseline methods. Notably, it outperforms the state-of-the-art defense method, HSIC-GAD, achieving +9.07% (84.22 vs. 75.15) on Cora and +20.07% (84.74 vs. 64.67) on Pubmed, averaged over different attacks. Due to the scalability limitations of existing attacks, we additionally evaluate AdaMF on the large-scale ogbn-arxiv dataset under clean and VoteAttack settings to further verify its effectiveness. Paired t-tests over 10 independent runs verify that AdaMF's improvements are statistically significant (p-value < 0.01). These results validate the effectiveness of adaptive spectral awareness and dual-channel contrastive fusion for adversarial defense.

To evaluate the defense performance of the proposed method under different perturbation strengths, we conduct experiments against the state-of-the-art attack method VoteAttack, where the feature perturbation strength is set to $\{0.2, 0.4, 0.6, 0.8, 1\}$. As shown in Figure 4, the performance of all baseline models decreases as the perturbation strength increases. However, AdaMF consistently exhibits a smaller performance degradation compared with other models.

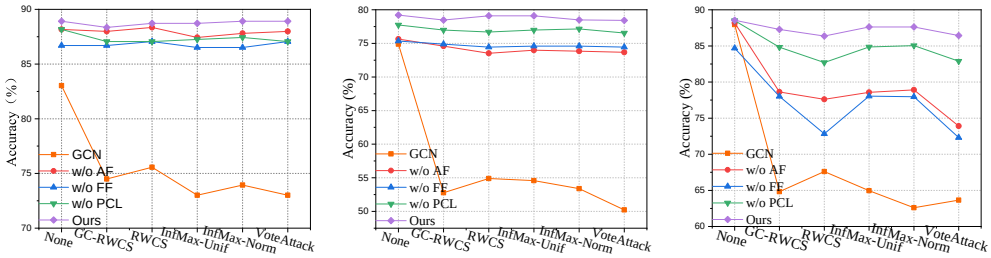
5.3 Ablation Study (RQ2)

To evaluate the effectiveness of each component in AdaMF, we construct three variants: (1) removing the adaptive filter (w/o AF); (2) removing the framelet filter (w/o FF); and (3) removing prototype contrastive learning (w/o PCL). Figure 5 demonstrates that introducing either adaptive filtering



(a) Cora (b) Citeseer (c) Pubmed

Figure 4: Results under different perturbation strengths.



(a) Cora (b) Citeseer (c) Pubmed

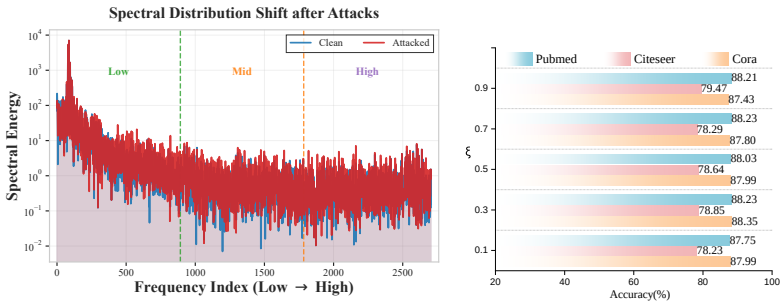
Figure 5: Ablation studies on various datasets.

or framelet filtering alone can improve the adversarial robustness of the model. Specifically, the adaptive filter dynamically adjusts the responses of different frequency bands, enabling the model to better adapt to attack-induced spectral variations. In contrast, the framelet filter leverages multi-resolution representations to capture localized perturbation patterns across different scales. When these two components are jointly incorporated, the model achieves further improvements, validating the complementary nature of adaptive frequency modeling and multi-scale representation learning. Furthermore, introducing prototype contrastive learning further enhances the consistency and discriminability of node representations, thereby improving the adversarial robustness of the model.

To further validate our motivation, we analyze the spectral variations of AdaMF under the same experimental settings. As shown in Figure 6(a), the proposed method maintains a stable spectral distribution before and after adversarial perturbations, demonstrating its effectiveness in mitigating attack-induced spectral variations.

5.4 Hyperparameter Sensitivity Analysis (RQ3)

We further investigate the impact of the prototype contrastive loss weight ξ on the performance of AdaMF, where ξ is set to $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. As shown in Figure 6(b), the model performance remains stable across different datasets with varying values of ξ, exhibiting only minor fluctuations. This indicates that AdaMF is insensitive to the selection of ξ and demonstrates strong robustness to this hyperparameter.



(a) (b)

Figure 6: (a) Spectral energy distributions of AdaMF under clean and adversarial settings; (b) Sensitivity analysis of the contrastive loss weight ξ.

6 Conclusion

In this paper, we propose AdaMF, a spectral-aware adversarially robust framework that integrates adaptive filters and framelet filters. The adaptive filter dynamically adjusts the low-, high-, and mid-frequency components to adapt to diverse spectral variations caused by adversarial perturbations applied to the input at test time. Furthermore, we introduce the framelet filter, which supports multi-scale graph signal analysis and enables the model to capture both global semantic information and local structural details. To further enhance the discriminability of representations under adversarial scenarios, we incorporate a prototype contrastive module. In future work, we will explore more efficient adaptive filtering strategies for large-scale graphs and investigate the generalization ability of AdaMF under more complex and evolving adversarial attack scenarios.

Ethics statement

This work does not involve human subjects, personal data collection, or privacy-sensitive information. We use publicly available benchmark datasets for evaluation. Our goal is to improve the robustness and reliability of graph neural networks under adversarial perturbations. We have considered potential misuse of the proposed techniques and believe that this work primarily contributes to improving the security and trustworthiness of machine learning systems.

Reproducibility statement

To ensure the reproducibility of the work and the facilitation of future research, we provide detailed descriptions of the experimental settings in Section 5.1 and Appendix D. The complete source code, together with processed datasets, is available in the Supplementary Material.

References

J Bruna, W Zaremba, A Szlam, and Y LeCun. Spectral networks and deep locally connected networks on graphs, 2nd int. In *Proceedings of the 2nd International Conference on Learning Representations*, 2014.

Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1115–1124. PMLR, 2018.

Bin Dong. Sparse representation on graphs by tight wavelet frames and applications. *Applied and Computational Harmonic Analysis*, 42(3):452–479, 2017.

Negin Entezari, Saba A Al-Sayouri, Amirali Darvishzadeh, and Evangelos E Papalexakis. All you need is low (rank) defending against adversarial attacks on graphs. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 169–177, 2020.

Jin Fan, Zheyu Wang, Zehao Wang, Jiajun Yang, Huifeng Wu, and Jia Wu. Enhancing gcn robustness against structural attacks via adaptive spectrum filtering. *IEEE Transactions on Information Forensics and Security*, 2025.

Yuxing Guo, Jianqing Liang, Kaixuan Yao, Zhihao Guo, and Jiye Liang. Graph adversarial defense via hilbert-schmidt independence criterion against influence maximization attacks. In *Proceedings of the ACM Web Conference 2026*, pp. 4565–4576, 2026.

Zhihao Guo, Ting Han, Peng Song, Chenjiao Feng, Kaixuan Yao, and Jiye Liang. Hyperbolic multicriteria rating recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2162–2171, 2025.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in Neural Information Processing Systems*, 33:22118–22133, 2020.

Jincheng Huang, Lun Du, Xu Chen, Qiang Fu, Shi Han, and Dongmei Zhang. Robust mid-pass filtering graph convolutional networks. In *Proceedings of the ACM Web Conference 2023*, pp. 328–338, 2023.

Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. Graph structure learning for robust graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 66–74, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2014.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*, 2016.

Enxia Li, Mengshi Chen, Sheng Xiang, and Ling Chen. Graph learning-empowered financial fraud detection: progress and future directions. *Intelligent Computing*, 4:0146, 2025.

Ming Li, Ruiting Zhao, Zihao Yan, Lu Bai, Lixin Cui, and Feilong Cao. Hyperaim: hypergraph contrastive learning with adaptive multi-frequency filters. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence*, volume 40, pp. 23063–23070, 2026.

Xiaorui Liu, Wei Jin, Yao Ma, Yaxin Li, Hua Liu, Yiqi Wang, Ming Yan, and Jiliang Tang. Elastic graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6837–6849. PMLR, 2021.

Jiaqi Ma, Shuangrui Ding, and Qiaozhu Mei. Towards more practical adversarial attacks on graph neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pp. 4756–4766, 2020.

Jiaqi Ma, Junwei Deng, and Qiaozhu Mei. Adversarial attack on graph neural networks as an influence maximization problem. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*, pp. 675–685. ACM, 2022.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bring order to the web. In *Proc. of the 7th International World Wide Web Conf.–1998*, 1999.

Lichao Sun, Yingtong Dou, Carl Yang, Kai Zhang, Ji Wang, S Yu Philip, Lifang He, and Bo Li. Adversarial attack and defense on graph data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):7693–7711, 2022.

Jacob Sweeten, Abdulrahman Takiddin, Muhammad Ismail, Shady S Refaat, and Rachad Atat. Cyber-physical gnn-based intrusion detection in smart power grids. In *2023 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pp. 1–6. IEEE, 2023.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, volume 36, pp. 8432–8440, 2022.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

Jihong Wang, Minnan Luo, Fnu Suya, Jundong Li, Zijiang Yang, and Qinghua Zheng. Scalable attack on graph data by injecting vicious nodes. *Data Mining and Knowledge Discovery*, 34(5): 1363–1389, 2020.

Liangliang Wen, Jiye Liang, Kaixuan Yao, and Zhiqiang Wang. Black-box adversarial attack on graph neural networks with node voting mechanism. *IEEE Transactions on Knowledge and Data Engineering*, 36(10):5025–5038, 2024.

Huijun Wu, Chen Wang, Yuriy Tyshetskiy, Andrew Docherty, Kai Lu, and Liming Zhu. Adversarial examples for graph data: Deep insights into attack and defense. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 4816–4823, 2019.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2020.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Beyond low-pass filtering: Graph convolutional networks with automatic filtering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):6687–6697, 2022.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5453–5462. PMLR, 2018.

Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 40–48. PMLR, 2016.

Hao Yu, Ke Liang, Dayu Hu, Wenxuan Tu, Chuan Ma, Sihang Zhou, and Xinwang Liu. Gzoo: black-box node injection attack on graph neural networks via zeroth-order optimization. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Xiang Zhang and Marinka Zitnik. Gnnguard: Defending graph neural networks against adversarial attacks. *Proceedings of the Advances in Neural Information Processing Systems*, 33:9263–9275, 2020.

Xuebin Zheng, Bingxin Zhou, Junbin Gao, Yuguang Wang, Pietro Lió, Ming Li, and Guido Montufar. How framelets enhance graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12761–12771. PMLR, 2021.

Dingyuan Zhu, Ziwei Zhang, Peng Cui, and Wenwu Zhu. Robust graph convolutional networks against adversarial attacks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1399–1407, 2019.

Daniel Zügner and Stephan Gunnemann. Adversarial attacks on graph neural networks via meta learning. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2847–2856, 2018.

A Theoretical Analysis of Spectral Response Adaptability and Perturbation Stability

We now characterize the spectral family actually searched by AdaMF and analyze its sensitivity to attack-induced graph perturbations. Let $q = 1 - \varepsilon$, $P = D^{-1/2}AD^{-1/2}$, and $L = I - P$, whose eigenvalues lie in $[0, 2]$. Since the hybrid filter is a polynomial in L (equivalently, in P), we write $F(L)$ interchangeably with $F(\hat{A})$ in what follows. Recall that the learnable coefficients satisfy $(\alpha, \beta, \gamma) \in \Delta_2$, where $\Delta_2 = \{(\alpha, \beta, \gamma) \in \mathbb{R}^3_+ : \alpha + \beta + \gamma = 1\}$.

Proposition 1 (Convex spectral-response family)**.** *The family $\mathcal{R}_\varepsilon$ is the convex hull $\mathrm{conv}\{g_L, g_H, g_M\}$; writing $r(\lambda) = c_0 + c_1\lambda + c_2\lambda^2$ for $r \in \mathcal{R}_\varepsilon$, its second-order coefficient satisfies $c_2 = -q^2\gamma$; moreover, every $r \in \mathcal{R}_\varepsilon$ satisfies $\max_{\lambda\in[0,2]} |r(\lambda)| \le 1$, and hence $\|\mathcal{F}_{\mathrm{Hybrid}}\|_2 \le 1$.*

Proof. Since $P = I - L$, we have $\mathcal{F}_{\mathrm{L}} = \varepsilon I + qP = I - qL$ and $\mathcal{F}_{\mathrm{H}} = \varepsilon I - qP = (2\varepsilon - 1)I + qL$, which yield g_L and g_H. For the mid-pass filter, completing the square and using $\varepsilon^2 - q^2 = 2\varepsilon - 1$ gives $\mathcal{F}_{\mathrm{M}} = \varepsilon^2 I - q^2(L - I)^2$, i.e., $g_M(\lambda) = \varepsilon^2 - q^2(\lambda - 1)^2$.

The first claim is the definition of a convex combination. For the second claim, substituting $\gamma = 1 - \alpha - \beta$ into $r = g_M + \alpha(g_L - g_M) + \beta(g_H - g_M)$ and comparing the coefficient of λ^2 on both sides gives $c_2 = -q^2\gamma$.

For the last claim, on $[0, 2]$ the linear responses g_L and g_H range between $2\varepsilon - 1 \in (-1, 1)$ and 1. The vertex form of g_M shows that its maximum is ε^2, attained at $\lambda = 1$, while its minimum on $[0, 2]$ is attained at the endpoints and equals $2\varepsilon - 1$. Hence $|g_L|, |g_H|, |g_M| \le 1$, and by convexity $|r(\lambda)| \le 1$ for every $r \in \mathcal{R}_\varepsilon$. The operator bound follows from $\|\mathcal{F}_{\mathrm{Hybrid}}\|_2 = \max_k |r(\lambda_k)| \le 1$. □

Proposition 1 shows that, under the simplex constraint, AdaMF searches a *bounded convex family* of spectral responses rather than the entire space $\mathcal{P}_2$ of quadratic polynomials. The role of each base filter within this family is characterized next.

Proposition 2 (Second-order flexibility induced by the mid-pass filter)**.** *Without the mid-pass filter, every convex combination of g_L and g_H is linear in λ, so the response set reduces to the one-dimensional edge $\mathrm{conv}\{g_L, g_H\} \subset \mathcal{P}_1$. With the mid-pass filter, $c_2 = -q^2\gamma$, so γ continuously controls the spectral curvature of the hybrid response. Finally, for the interpolating family $\alpha = \beta = (1 - \gamma)/2$, $\gamma \in [0, 1]$, the hybrid response admits the closed form*

$$r_\gamma(\lambda) = (1 - \gamma)\varepsilon + \gamma g_M(\lambda) = (1 - \gamma)\varepsilon + \gamma\varepsilon^2 - \gamma q^2(\lambda - 1)^2 \tag{12}$$

and satisfies

$$r'_\gamma(1) = 0, \qquad \mathcal{J}(r_\gamma) := \max_{\lambda\in[0,2]} |r'_\gamma(\lambda)| = 2q^2\gamma. \tag{13}$$

Hence γ controls the trade-off between spectral smoothness and frequency selectivity: $\gamma = 0$ recovers the flat response ε, while increasing γ restores second-order spectral structure at the cost of a linearly growing spectral sensitivity.

Proof. For the first claim, any $r = \alpha g_L + (1 - \alpha)g_H$ with $\alpha \in [0, 1]$ satisfies

$$r(\lambda) = \alpha + (1 - \alpha)(2\varepsilon - 1) + q(1 - 2\alpha)\lambda, \tag{14}$$

which is linear in λ, and the set $\{\alpha g_L + (1 - \alpha)g_H : \alpha \in [0, 1]\}$ is exactly the edge of the simplex connecting g_L and g_H. The second claim restates Proposition 1.

For the last claim, since $g_L(\lambda) + g_H(\lambda) = 2\varepsilon$, taking $\alpha = \beta = (1 - \gamma)/2$ gives $r_\gamma(\lambda) = (1 - \gamma)\varepsilon + \gamma g_M(\lambda) = (1 - \gamma)\varepsilon + \gamma\varepsilon^2 - \gamma q^2(\lambda - 1)^2$, where the last equality uses the vertex form of g_M. Differentiating, $r'_\gamma(\lambda) = -2\gamma q^2(\lambda - 1)$, so $r'_\gamma(1) = 0$ and $\mathcal{J}(r_\gamma) = \max_{\lambda\in[0,2]} 2\gamma q^2|\lambda - 1| = 2q^2\gamma$. □

Proposition 2 clarifies the division of roles among the three filters: the low-pass and high-pass filters provide first-order spectral adjustments with opposite slopes, while the mid-pass filter is the only component that introduces a nonzero second-order coefficient $c_2 = -q^2\gamma$, thereby lifting the

response space from the edge $\mathrm{conv}\{g_L, g_H\} \subset \mathcal{P}_1$ to the full convex hull $\mathcal{R}_\varepsilon$. Theorem 1 and Theorem 2 quantify how this structure affects response stability under spectral perturbations, and their proofs are provided as follows.

Proof. The first inequality follows from the mean value theorem. For r_γ, its vertex form gives, for any admissible λ, δ_λ with $\lambda + \delta_\lambda \in [0, 2]$,

$$\begin{aligned} r_\gamma(\lambda + \delta_\lambda) - r_\gamma(\lambda) &= -\gamma q^2 \big[(\lambda + \delta_\lambda - 1)^2 - (\lambda - 1)^2\big] \\ &= -\gamma q^2 \big[2(\lambda - 1)\delta_\lambda + \delta_\lambda^2\big] \\ &= 2q^2\gamma(1 - \lambda)\delta_\lambda - \gamma q^2 \delta_\lambda^2. \end{aligned} \tag{15}$$

Setting $\lambda = 1$ yields the second-order identity. □

Theorem 1 characterizes the spectral sensitivity of the proposed filters at the scalar level. The following result extends the analysis from eigenvalue perturbations to graph perturbations, so that the assumption on spectral drift is grounded in the graph distance between the clean and attacked Laplacians.

Proof. For the first statement, the operator forms $\mathcal{F}_L = I - qL$, $\mathcal{F}_H = (2\varepsilon - 1)I + qL$, and $\mathcal{F}_M = \varepsilon^2 I - q^2(L - I)^2$ give

$$\|\mathcal{F}_L(\widetilde{L}) - \mathcal{F}_L(L)\|_2 \le q\eta, \qquad \|\mathcal{F}_H(\widetilde{L}) - \mathcal{F}_H(L)\|_2 \le q\eta, \tag{16}$$

and

$$\begin{aligned} \|\mathcal{F}_M(\widetilde{L}) - \mathcal{F}_M(L)\|_2 &= q^2 \big\|(\widetilde{L} - I)\Delta + \Delta(L - I)\big\|_2 \\ &\le q^2 \big(\|\widetilde{L} - I\|_2 + \|L - I\|_2\big)\eta \le 2q^2\eta, \end{aligned} \tag{17}$$

where the last inequality uses $\|L - I\|_2 \le 1$ and $\|\widetilde{L} - I\|_2 \le 1$, since the eigenvalues of a normalized graph Laplacian lie in $[0, 2]$. The claimed bound follows by taking the convex combination with weights (α, β, γ).

For the last statement, the constant terms in $r_\gamma(L) = (1 - \gamma)\varepsilon I + \gamma\varepsilon^2 I - \gamma q^2(L - I)^2$ cancel in the difference, so

$$\|r_\gamma(\widetilde{L}) - r_\gamma(L)\|_2 \le \gamma q^2 \big(\|\widetilde{L} - I\|_2 + \|L - I\|_2\big)\eta \le 2q^2\gamma\,\eta. \tag{18}$$

□

Remark Unrolling the bound in Theorem 2 over T layers gives the recursion $\Delta_{t+1} \le a_t \Delta_t + b_t$ with $a_t = L_\sigma \|W_t\|_2$ and $b_t = 2q^2\gamma\eta L_\sigma \|W_t\|_2 \|\widetilde{Z}_t\|_F$, where $\Delta_t = \|\widetilde{Z}_t - Z_t\|_F$, which yields an explicit multi-layer sensitivity bound by induction.

Taken together, Propositions 1–2 and Theorems 1–2 provide a mechanism-level explanation of the adaptive multi-frequency design. The low-pass and high-pass filters contribute opposite first-order slopes whose cancellation yields the flat baseline ε, while the mid-pass filter introduces the sole second-order degree of freedom $c_2 = -q^2\gamma$, allowing the model to remain locally first-order insensitive at the center frequency $\lambda = 1$ while preserving discriminative spectral structure elsewhere. Accordingly, the hybrid filter adaptively trades spectral smoothness against frequency selectivity within a bounded and perturbation-characterized response family.

We emphasize that Theorems 1 and 2 characterize the perturbation sensitivity of the frequency-filtering branch at the representation level. They do not by themselves yield a classification-accuracy guarantee for the full AdaMF model, which also involves framelet filtering, feature fusion, and the supervised and contrastive objectives.

B Comparison of Spectral Modeling Strategies

To further demonstrate the differences between the proposed AdaMF and existing spectral defense methods in terms of frequency modeling strategies, we compare them from four perspectives: Single Spectral Response (SSR), Fixed Spectral Adaptation (FSA), Learnable Spectral Adaptation (LSA), and Multi-Scale Analysis (MSA). The detailed comparison results are presented in Table 2.

Table 2: Comparison of Frequency Modeling Capabilities

Method	SSR	FSA	LSA	MSA
GCN-SVD Entezari et al. (2020)	✓	✗	✗	✗
AutoGCN Wu et al. (2022)	✗	✗	✓	✗
Mid-GCN Huang et al. (2023)	✓	✗	✗	✗
ASF-GCN Fan et al. (2025)	✗	✓	✗	✗
AdaMF (Ours)	✗	✗	✓	✓

C PSEUDOCODE

To provide a clearer understanding of the overall learning procedure of AdaMF, we present the detailed pseudocode, as detailed in Algorithm 1.

Algorithm 1 AdaMF

Input: Graph $\mathcal{G}$, adjacency matrix $\mathbf{A}$, feature matrix $\mathbf{X}$
Output: Model $f_{\boldsymbol{\theta}}$ and representation $\mathbf{Z}$

```
Construct augmented views G_f and G_t via feature and topology perturbations.
Calculate the low-pass and high-pass coefficients W_j^0, W_j^i of the Framelet filters via Eq. 4.
while not converged do
    Compute adaptive spectral representations for each view via Eq. 5.
    Compute Framelet-filtered representations for each view via Eq. 7.
    Compute the loss via Eq. 11.
end while
return θ and Z
```

D ADDITIONAL EXPERIMENT DETAILS AND RESULTS

In this section, we provide a detailed description of the benchmark datasets, baseline methods, and attack strategies employed in this paper, and present the visualization analysis of the proposed AdaMF.

D.0.1 BENCHMARK DATASETS

The datasets used in this study, including Cora Yang et al. (2016), Citeseer, Pubmed, and ogbn-arxiv Hu et al. (2020), are widely adopted benchmark datasets. Detailed statistics of these datasets are presented in Table 3.

Table 3: Statistics of the datasets

Dataset	Nodes	Edges	Classes	Features
Cora	2,708	5,429	7	1,433
Citeseer	3,327	4,732	6	3,703
Pubmed	19,717	44,338	3	500
ogbn-arxiv	169,343	2,484,941	40	128

D.0.2 BASELINE MODELS

To evaluate the adversarial robustness of the proposed method, we compare it with ten baselines, which are grouped into three categories: (1) classical GNNs, including GCN Kipf & Welling (2016), GAT Veličković et al. (2017), and JKNet Xu et al. (2018); (2) spatial-domain defense methods, including RGCN Zhu et al. (2019), Elastic Liu et al. (2021), and GAD Guo et al. (2026); and (3) spectral-domain defense methods, including SVD Entezari et al. (2020), AutoGCN Wu et al. (2022), Mid-GCN Huang et al. (2023) and ASF-GCN Fan et al. (2025). The detailed descriptions of these baselines are presented as follows:

Classical GNNs

- **GCN**: This method proposes an efficient layer-wise propagation algorithm based on the first-order approximation of spectral convolution.
- **GAT**: This model leverages a masked self-attention mechanism to assign different weights to neighboring nodes and learn effective representations.
- **JKNet**: This method introduces a novel node representation aggregation scheme that retains representations from different propagation layers and aggregates multi-layer information through jumping connections, enabling node-wise adaptive neighborhood selection.

Spatial-domain Defense Methods

- **RGCN**: This model represents each node as a Gaussian distribution, with the mean capturing the node representation and the variance characterizing its uncertainty. The learned variance is then leveraged to assign adaptive weights to edges for information aggregation.
- **Elastic**: Elastic leverages L_1 and L_2 graph smoothness regularization to enable adaptive smoothing in GNNs, and introduces an elastic message-passing mechanism accordingly.
- **GAD**: This method, abbreviated as HSIC-GAD, is denoted as GAD due to space limitations. It improves adversarial robustness by introducing Hilbert-Schmidt independence criterion regularization to constrain the correlations between different layers.

Spectral-domain Defense Methods

- **GCN-SVD**: Based on the insight that adversarial attacks mainly influence high-rank components, this method applies low-rank approximation to the data, aiming to reduce adversarial noise perturbations.
- **AutoGCN**: This method proposes an automatic convolutional network that dynamically learns and adjusts the filter shapes in the spectral domain.
- **Mid**: The full name of this method is Mid-GCN, which is denoted as Mid due to space limitations. It investigates the impact of middle-frequency filters on adversarial robustness and accordingly designs a middle-frequency convolutional network.
- **ASF-GCN**: This method enhances the inherent robustness of GCN by proposing a spectral adaptive model, which effectively improves the model's adversarial robustness.

D.0.3 Adversarial Attack Methods

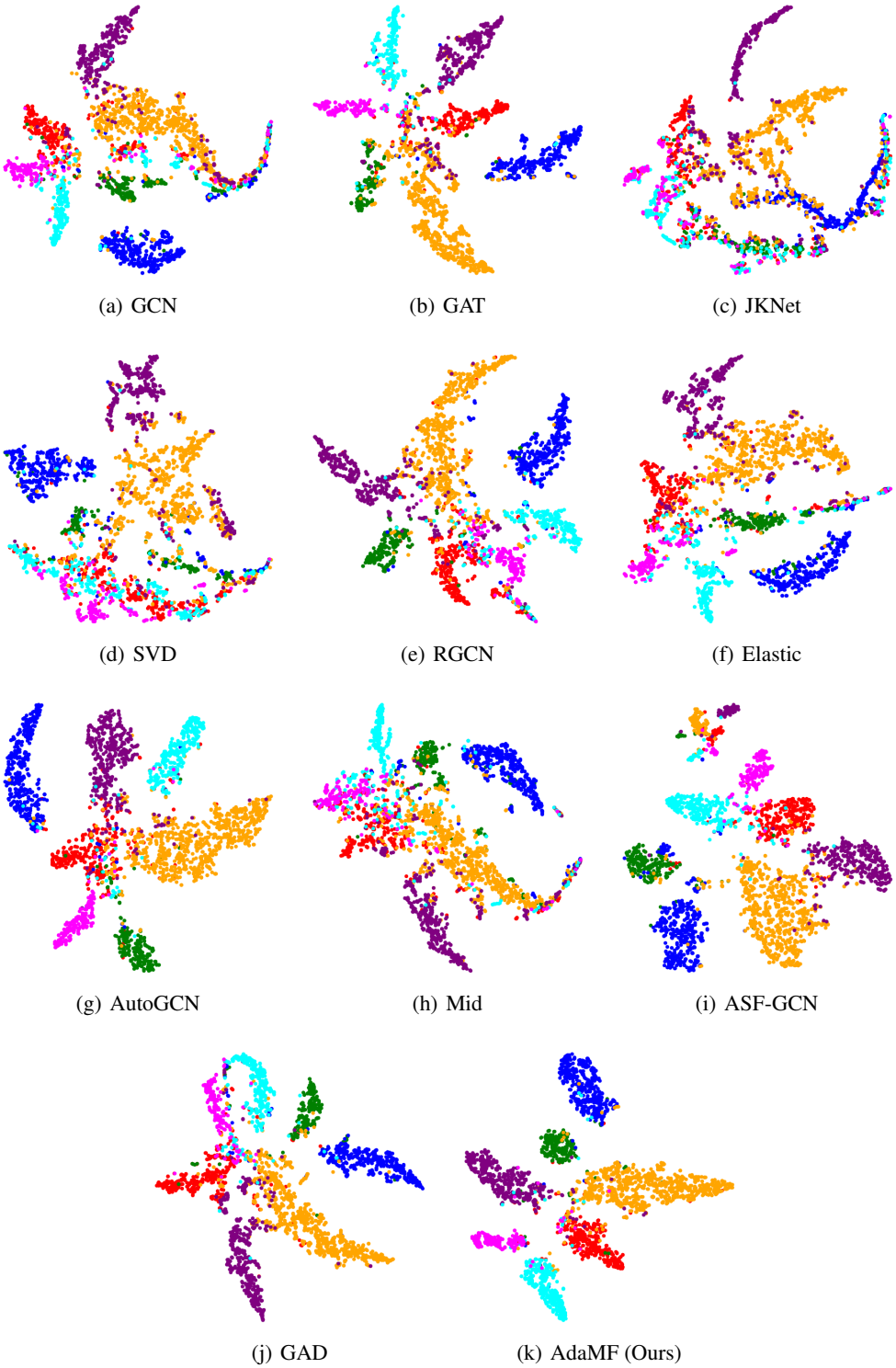
We categorize the adopted evasion attack methods into two categories: graph modification attacks, including GC-RWCS Ma et al. (2020), RWCS Ma et al. (2020), InfMax-Norm Ma et al. (2022), InfMax-Unif Ma et al. (2022), and VoteAttack Wen et al. (2024), and graph injection attacks, including AFGSM Wang et al. (2020) and GZOO Yu et al. (2024). Detailed descriptions of these methods are provided as follows:

Graph Modification Attacks

- **RWCS**: This method employs a random walk matrix with a restart mechanism to identify the optimal set of target nodes and utilizes domain knowledge to perturb the features of the selected nodes.
- **GC-RWCS**: Based on RWCS, this method incorporates an additional correction mechanism into the node selection process, enabling the selected nodes to be more representative.
- **InfMax-Norm, InfMax-Unif**: These methods formulate adversarial attacks as an influence maximization problem under the general threshold model and assigns node thresholds based on either a normal distribution or a uniform distribution.
- **VoteAttack**: This method establishes a connection between adversarial attacks and the voting matrix, based on which the attack node set is selected.

Graph Injection Attacks

- **AFGSM**: This method utilizes an approximate fast gradient sign method to inject new nodes and edges into the original graph, thereby generating adversarial perturbations.



(a) GCN (b) GAT (c) JKNet

(d) SVD (e) RGCN (f) Elastic

(g) AutoGCN (h) Mid (i) ASF-GCN

(j) GAD (k) AdaMF (Ours)

Figure 7: t-SNE visualizations of node representations under the VoteAttack scenario on the Cora dataset.

- **GZOO**: This method employs zero-order optimization to search for the optimal injected nodes and features, eliminating the need for training a surrogate model.

Table 4: Performance under extreme perturbation budgets

Dataset	Cora	Citeseer	Pubmed
SVD	21.39±6.86	31.65±3.86	37.83±6.90
RGCN	28.72±6.48	33.56±8.60	52.56±11.28
Elastic	25.58±7.16	29.34±5.56	37.88±10.15
AutoGCN	64.68±9.33	50.89±9.38	39.68±8.10
Mid-GCN	36.19±4.40	33.64±5.48	43.85±5.42
ASF-GCN	58.52±4.42	40.20±7.26	46.89±7.29
GAD	62.03±12.55	52.67±7.85	52.44±15.64
Ours	**67.87±8.31**	**53.73±6.98**	**57.27±6.34**

D.1 Performance under Extreme Attack Scenarios

To evaluate the adversarial robustness of the model under extreme attack scenarios, we select 10% of the nodes in each dataset as attack targets under VoteAttack, which is ten times larger than the commonly adopted attack ratio (1%). The mean and standard deviation are reported over 10 independent runs, and the results are reported in Table 4. The experiments clearly show that our proposed method still achieves the best adversarial robustness under extreme settings. These results further demonstrate the effectiveness of the adaptive filters and framelet filters in enhancing adversarially robust representation learning.

D.2 Visualization of Robust Node Representations

In this section, we visualize the node representations of different baseline methods under the VoteAttack scenario on the Cora dataset. In the t-SNE visualization, each color represents one class of nodes. Clearer separations among clusters of different colors and fewer overlaps among different classes indicate more discriminative node representations. As shown in Figure 7, the node representations learned by AdaMF exhibit larger inter-class margins and clearer class boundaries compared with baseline methods. These results further demonstrate the effectiveness of the adaptive filtering and multi-scale filtering components in AdaMF for learning discriminative representations under adversarial attacks, thereby enhancing adversarial robustness.

AI use statement

In this work, we used generative AI tools for translation assistance. We have not used generative AI tools for interpreting results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for editing the manuscript to improve readability. We have reviewed all AI-assisted work. We used LLMs to check the grammar, word choice, punctuation, capitalization, and other linguistic aspects of the manuscript, and manually reviewed the modification suggestions generated by the LLMs to ensure their appropriateness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.