Cooperative Graph Transformer with Structural Consensus for Multi-View Learning

Zhiyuan Lai[1], **Jiacheng Li**[1], **Jiayuan Wang**[1], **Shiping Wang**[1*]

[1]College of Computer and Data Science, Fuzhou University, Fuzhou, China
zhiyuanlai001@163.com, jiacheng63@outlook.com, wwwangjiayuan@gmail.com, shipingwangphd@163.com

Abstract

Multi-view learning aims to effectively integrate data from different sources by exploring the consistency and complementarity across views. Current multi-view methods based on Graph Convolutional Networks (GCNs) primarily focus on local information, making it difficult to capture global dependencies. Furthermore, multi-view data typically lack explicit structural representations, and the topologies constructed via node similarity in existing approaches are prone to noise, while simple fusion strategies are often inadequate for effectively suppressing this noise and for uncovering meaningful structural information. To tackle these issues, this paper proposes CoGFormer, a cooperative graph transformer with structural consensus learning. CoGFormer maps multi-view data into a unified space and jointly models local and global consensus: a denoising structural consensus graph convolutional network refines the consensus graph to enhance local consistency and robustness, while a structure-guided attention mechanism explicitly injects high-order cross-view structural biases to capture global consistency and improve semantic coherence. Experiments on multiple benchmarks demonstrate that CoGFormer outperforms existing state-of-the-art methods, validating its effectiveness.

Introduction

Improvements in multimedia hardware and sensor capability have made multi-view observation of objects broadly feasible. Compared to single-view data, multi-view data offers more comprehensive and complementary feature representations. Consequently, a learning paradigm known as multi-view learning has emerged and achieved remarkable success in various areas, including social networks (Zhou et al. 2024), recommender systems (Cheng et al. 2022; Yu et al. 2023), and computer vision (Yang et al. 2023; Li et al. 2023).

In recent years, driven by advances in Graph Convolutional Networks (GCNs) (Kipf and Welling 2017), an increasing number of studies have focused on integrating graph-based models into multi-view learning frameworks (Zhao et al. 2025; Kong et al. 2025). However, current investigations into GCN-based multi-view semi-supervised classification primarily concentrate on two directions: 1) constructing intra-view adjacency graphs to extract local structural representations (Wang et al. 2024; Li, Ni, and Wang 2024; Wen et al. 2024), and 2) leveraging cross-view alignment graphs to enable information propagation and sharing (Lou et al. 2024; Zheng et al. 2022). Notwithstanding these efforts, most GCN-based methods remain primarily limited in their ability to model global dependencies, often neglecting the need for deeper structural context beyond local neighborhoods.

To address the limitations of traditional GCNs in modeling global dependencies, Transformers (Vaswani et al. 2017) have emerged as a foundational encoder framework for graph-structured data, finding widespread application across diverse domains (Wu et al. 2023a; Chen, O'Bray, and Borgwardt 2022). By treating graph nodes as input tokens and utilizing their global self-attention mechanism to compute pairwise attention scores, Transformers effectively capture complex global dependencies. Nevertheless, the self-attention mechanism's reliance on non-fixed structural priors poses challenges in adaptability when processing the heterogeneous and multi-faceted structural information inherent in multi-view data, potentially limiting its effectiveness in such scenarios. Moreover, multi-view data inherently lacks an explicit graph structure. Existing GCNs-based multi-view methods often construct topological structures by computing the similarity of geometric relationships between nodes. This implies that the topological structure morphology of each view is adjustable. When constructing complex topological structures with densely connected edges for each view, while potentially enriching the view with abundant structural information, it also introduces substantial noise due to errors in similarity computation or structural inconsistency across views. In such scenarios, relying solely on simple averaging or weighted fusion of structural features from each view not only struggles to effectively suppress noise but also fails to fully exploit latent structural information. Therefore, two key challenges remain to be addressed: 1) how to integrate GCNs and attention mechanisms to effectively leverage structural information in multi-view data for extracting both local and global representations, and 2) how to fully utilize the latent structural consistency in multi-view settings to mitigate noise interference and improve the robustness of the overall representation.

To tackle these challenges, we propose a novel multi-

*Corresponding Authors.

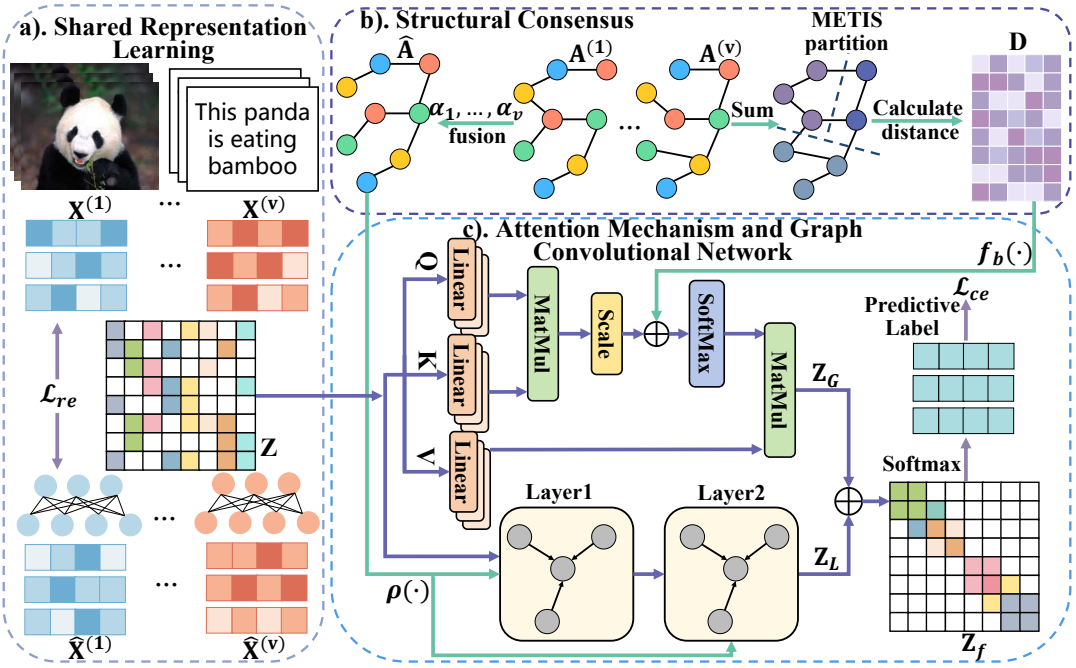


Figure 1: Overview of the proposed CoGFormer: a) Multi-view features are projected into a unified latent space to learn a shared representation $\mathbf{Z}$; b) View-specific adjacency matrices are weighted and summed to form a consensus graph $\hat{\mathbf{A}}$ and a METIS-input graph for deriving the community-level distance matrix $\mathbf{D}$; c) $\hat{\mathbf{A}}$ is denoised by a shrinkage function $\rho(\cdot)$ for a GCN to extract $\mathbf{Z}_L$, while $\mathbf{D}$ is mapped by $f_b(\cdot)$ into bias terms for attention mechanism to obtain $\mathbf{Z}_G$; finally, $\mathbf{Z}_L$ and $\mathbf{Z}_G$ are fused into the representation $\mathbf{Z}_f$.

view learning model, Cooperative Graph Transformer with Structural Consensus Learning (CoGFormer). Specifically, we first map multi-view information into a unified latent space to obtain an initial inter-view consensus representation. Based on this representation, we design a differentiable function to construct a structural consensus graph, which aims to filter out noisy edges caused by view inconsistencies or low-quality data and to enhance local structural consensus. We further utilize a graph convolutional network to mine local consensus patterns. To capture cross-view global structural consistency, we introduce high-order structural bias terms into the multi-head attention mechanism, enabling the model to explicitly learn global dependencies spanning multiple views. By jointly optimizing these modules, CoGFormer reinforces local structural consensus while maintaining global semantic coherence, resulting in a robust and consistent fused representation. The framework of the proposed CoGFormer is shown in Figure 1. Our main contributions are summarized as follows:

- We propose a novel joint learning framework that integrates a denoised structural consensus graph convolution with a structure-guided attention mechanism, leveraging structural information to effectively extract both local and global representations.
- We optimize multi-view local representation by designing a differentiable function to eliminate noisy edges on fused graphs, while explicitly capturing global associations by incorporating higher-order bias terms enriched with cross-view structural information into the attention mechanism.
- Extensive experiments on multi-view semi-supervised classification tasks demonstrate that our approach outperforms state-of-the-art methods, showcasing superior effectiveness and robustness.

Related Work

Graph-based Multi-view Learning

Graph-based multi-view learning has become a powerful paradigm for integrating multi-source information by uncovering both shared and complementary relationships among views. Prior studies have developed various approaches for multi-view feature fusion and robustness enhancement. For example, (Cheng et al. 2020) introduced a dual-path encoder to jointly learn graph embeddings and view-consistent features. (Tang et al. 2020) proposed a parameter-free cross-view graph diffusion method that iteratively refined view-specific graphs for clustering. (Liang et al. 2022) developed a min-max robust learning formulation to resist noisy views without introducing extra hyperparameters. (Zhou and Du 2023) designed a consensus graph filter learning strategy to improve multi-view clustering. (Pu et al. 2024) addressed incomplete views using view-specific encoders, partial graph construction, and adaptive imputation. Meanwhile, (Xiong et al. 2025) integrated contrastive multi-view learning with graph-based social recommendation to better model user preferences. While these methods

have advanced multi-view learning, a fundamental challenge remains: how to reliably construct a high-quality consensus graph that is robust to view-specific noise and inconsistencies. Many existing approaches either implicitly assume clean input graphs or lack explicit mechanisms to denoise and refine the fused structure—potentially propagating errors in downstream tasks. Our approach directly addresses this issue by building a denoised consensus graph, providing a robust foundation for learning both local and global representations.

Graph Transformer

Graph Neural Networks (GNNs), leveraging their efficient modeling capabilities for graph-structured data, complement Transformers in their strengths of capturing long-range dependencies and representation learning. This synergy has spurred Graph Transformers (GTs) into prominence as a key research focus, yielding notable advancements across diverse tasks. For instance, (Peng et al. 2023) extended GTs into collaborative learning by integrating graph smoothing and multi-modal embeddings. (Zhang et al. 2022) treated node sampling as a combinatorial bandit problem and introduced a hierarchical attention mechanism with graph coarsening to reduce computational cost. (Liu et al. 2023) combined GTs with graph pooling to coarsen large-scale graphs while preserving long-range dependencies. (Wang et al. 2025) designed a multi-modal GT that deeply fused structural and attribute information. (Zhang et al. 2025) proposed a topology-aware GT by injecting topological features and leveraging mutual information to enhance link prediction. (Gui, Ye, and Xiao 2024) introduced a structure-aware GT to mitigate performance loss caused by feature distribution shifts during parameter-efficient fine-tuning. Despite these advances, existing GTs are predominantly designed for single-view graphs, limiting their applicability in multi-view settings. In such scenarios, a key challenge is to integrate heterogeneous and potentially conflicting structures from multiple sources. The standard self-attention mechanism in GTs—relying solely on node features—may incorrectly connect unrelated nodes across views due to the absence of a cross-view structural prior. Thus, developing GT frameworks capable of effectively integrating and denoising multi-view graph structures remains an important open problem.

The Proposed Method

In this section, we first present the three core components of CoGFormer, namely shared representation learning, denoised structural consensus graph convolutional network and structure-guided attention mechanism, followed by the training details. For clarity, the mathematical notations used throughout the paper are summarized in Table 1.

Notations	Descriptions
$\mathbf{X}^{(v)} \in \mathbb{R}^{N\times d_v}$	The v-th view feature matrix.
$\mathbf{A}_c \in \mathbb{R}^{C\times C}$	The community adjacency matrix.
$\mathbf{A}^{(v)} \in \mathbb{R}^{N\times N}$	The v-th view adjacency matrix from $\mathbf{X}^{(v)}$.
$\mathbf{Z}_p \in \mathbb{R}^{C\times d}$	The low-rank projection matrix.
$\mathbf{Z}_f \in \mathbb{R}^{N\times c}$	The final representation.
$\mathbf{T} \in \mathbb{R}^{N\times N}$	The data-driven threshold matrix.
$\mathbf{D} \in \mathbb{R}^{N\times C}$	The distance matrix.
$\rho(\cdot)$	Differentiable shrinkage enhanced function .
ϕ	The community assignment function.
α	Learnable parameter of the fusion matrix.

Table 1: The symbolic notations and their descriptions.

Shared Representation Learning

To effectively learn a unified representation from multi-view data, we employ an encoder-decoder reconstruction framework. Let $\mathbf{X} = \{\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(V)}\}$ denote the multi-view data, where $\mathbf{X}^{(v)} \in \mathbb{R}^{N\times d_v}$ represents the feature matrix of the v-th view (N denotes the number of nodes, and d_v is the feature dimension of the v-th view). The shared representation $\mathbf{Z} \in \mathbb{R}^{N\times d}$ is then learned via an multi-layer encoder,

$$\mathbf{Z}^{(l)} = \sigma(\mathbf{Z}^{(l-1)}\mathbf{W}_e^{(l)}), \tag{1}$$

where $\mathbf{Z}^{(0)} \in \mathbb{R}^{N\times N}$ is initialized as the identity matrix, σ is the ReLU activation function to ensure non-negativity, and $\mathbf{W}_e^l$ denotes the weight of the encoder at layer l.

Subsequently, the shared representation $\mathbf{Z}$ is reconstructed into the original feature space of each view through the decoder,

$$\hat{\mathbf{X}}^{(v)} = \sigma(\sigma(\sigma(\mathbf{Z}\mathbf{W}_d^{(v,1)})\cdots)\mathbf{W}_d^{(v,l)}), \tag{2}$$

where $\mathbf{W}_d^{(v,l)}$ denotes the decoder weight of the l-th layer for the v-th view.

Finally, to ensure that the obtained shared representation $\mathbf{Z}$ captures consistent information from each view, we employ a reconstruction loss to evaluate the quality of $\mathbf{Z}$,

$$\mathcal{L}_{re} = \frac{1}{2}\sum_{v=1}^{V}\|\mathbf{X}^{(v)} - \hat{\mathbf{X}}^{(v)}\|_F^2, \tag{3}$$

where $\hat{\mathbf{X}}^{(v)} \in \mathbb{R}^{N\times d_v}$ denotes the reconstructed original features of the v-th view from the shared representation $\mathbf{Z}$, and $\|.\|_F$ represents the Frobenius norm.

Denoised Structural Consensus Graph Convolutional Network

The learning of the shared representation $\mathbf{Z}$ does not take into account structural information. To address this, we introduce Graph Convolutional Network (GCN) to integrate multi-view topological structures. Specifically, we first fuse the original topological structures from different views,

$$\begin{aligned} \alpha^{(v)} &\leftarrow \frac{\exp(\alpha^{(v)})}{\sum_{v=1}^{V}\exp(\alpha^{(v)})}, \\ \hat{\mathbf{A}} &= \sum_{v=1}^{V}\alpha^{(v)}\mathbf{A}^{(v)}, \end{aligned} \tag{4}$$

where each $\alpha^{(v)}$ is a learnable parameter. These parameters are constrained by $\sum_{v=1}^{V}\alpha^{(v)} = 1$ and are re-normalized

using the softmax at each epoch. Here, $\mathbf{A}^{(v)}$ denotes the initialized adjacency matrix for the v-th view.

Due to the inconsistent data quality across different views, the generated adjacency matrix inevitably contains noisy edges. Weighted fusion alone may fail to effectively suppress noise; thus, we introduce the Differentiable Shrinkage Enhancement (DSE) function to optimize the fused adjacency matrix, thereby obtaining a denoised structural consensus graph. This function $\rho(\cdot)$ can be defined as follows:

$$\begin{aligned} \mathbf{S} &= \text{Sigmoid}\left(\delta\left(\hat{\mathbf{A}} - \mathbf{T}\right)\right), \\ \mathbf{E} &= \left[\mathbf{1} + \text{ReLU}\left(\hat{\mathbf{A}} - \mathbf{T}\right)\right], \\ \rho(\hat{\mathbf{A}}) &= \hat{\mathbf{A}} \odot \mathbf{S} \odot \mathbf{E}, \end{aligned} \tag{5}$$

where $\mathbf{T}$ denotes the data-driven threshold matrix, $\mathbf{S}$ and $\mathbf{E}$ denote the data shrinkage matrix and the data enhancement matrix, respectively. Here, $\odot$ denotes hadamard product, and δ denotes the steepness coefficient of the shrinkage function, when $\hat{A}_{ij} < T_{ij}$, setting $\hat{A}_{ij} \approx \hat{A}_{ij} \cdot \text{Sigmoid}(\delta\Delta)$ corresponds to exponential attenuation, whereas when $\hat{A}_{ij} > T_{ij}$, setting $\hat{A}_{ij} \approx \hat{A}_{ij} \cdot (1 + \Delta)$ reflects linear enhancement, where $\Delta = \hat{A}_{ij} - T_{ij}$ denotes the deviation from the data-driven threshold. The design of the data-driven threshold matrix $\mathbf{T} \in \mathbb{R}^{N \times N}$ is presented as follows:

$$\mathbf{T} = \text{Sigmoid}(\frac{1}{r}\mathbf{Z}\mathbf{W}_t(\mathbf{Z}\mathbf{W}_t)^T + \theta\mathbf{1}\mathbf{1}^T), \tag{6}$$

where r is the projection dimension, $\mathbf{W}_t \in \mathbb{R}^{d \times r}$ is a learnable weight matrix, $\theta \in \mathbb{R}$ is a learnable global threshold parameter, and $\mathbf{1} \in \mathbb{R}^N$ is an all-ones vector. This formulation allows $\mathbf{T}$ to be data-dependent, balancing pairwise node interactions against a learnable global threshold. The propagation rule for our improved GCN is therefore given by:

$$\mathbf{Z}^{(l+1)} = \sigma(\hat{\mathbf{D}}^{-\frac{1}{2}}\rho(\hat{\mathbf{A}})\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{Z}^{(l)}\mathbf{W}_{gcn}^{(l)}), \tag{7}$$

where σ is the activation function, $\mathbf{W}_{gcn}^{(l)}$ is the learnable weight matrix of layer l, and $\hat{\mathbf{D}}$ is the degree matrix of $\rho(\hat{\mathbf{A}})$. Propagating $\mathbf{Z}$ through the GCN yields the local representation $\mathbf{Z}_L$, which captures structural consensus.

Structure-guided Attention Mechanism

To address the limitation of GCN in effectively capturing the global receptive field in complex multi-view learning scenarios, we combine the attention mechanism of the Transformer with the GCN. However, the attention mechanism lacks structural information to guide this process. To effectively leverage multi-view structural information, we first fuse the adjacency matrices $\mathbf{A}^{(v)}$ of all views to facilitate subsequent processing:

$$\mathbf{A}_f = \text{sign}\left(\sum_{v=1}^{V}\mathbb{I}\left(\mathbf{A}^{(v)} \neq 0\right)\right), \tag{8}$$

where $\mathbb{I}$ denotes the indicator function, and $\text{sign}(\cdot)$ represents the sign function, which facilitates the implementation of binarization. Then, we employ the METIS algorithm (Karypis and Kumar 1998) to partition the graph structure into C balanced communities, defining a community assignment function $\phi : N \to 1, \cdots, C$. This algorithm ensures high-quality community structures by effectively balancing intra-community connectivity and inter-community sparsity. From these assignments, we leverage the edge set E of $\mathbf{A}_f$ to build a community adjacency matrix $\mathbf{A}_c \in \{0,1\}^{C \times C}$, where a connection between communities p and q is indicated by $\mathbf{A}_c(p,q) = 1$.

To further explore the high-order structural consensus information of the graph, we define a distance matrix $\mathbf{D} \in \mathbb{R}^{N \times C}$, which is computed based on the shortest-path distances in $\mathbf{A}_c$. The details are as follows:

$$D_{ic} = \begin{cases} \min(d_{A_c}(\phi(i), c), \tau), & \text{if } c \text{ reachable} \\ \tau, & \text{otherwise} \end{cases}, \tag{9}$$

where $d_{A_c}(p,c)$ denotes the shortest-path length between communities p and c in the community adjacency graph, and τ is a saturation threshold that prevents overly large distances.

Since the computational complexity of the traditional Transformer's attention mechanism is $\mathcal{O}(N^2)$, to reduce the time cost, inspired by (Luo et al. 2024), we project $\mathbf{Z}$ using a matrix $\mathbf{P}$, as detailed below:

$$\mathbf{Z}_p = (\mathbf{P}^T\mathbf{Z}) \oslash (\mathbf{P}^T\mathbf{1}), \tag{10}$$

where $\oslash$ denotes the element-wise division, $\mathbf{P} = \text{OneHot}(\phi) \in \{0,1\}^{N \times C}$ represents the community assignment matrix, where each row represents the community membership of the corresponding node, specifically, where the entry P_{ic} is 1 if $\phi(i) = c$ and 0 otherwise, and $\mathbf{Z}_p \in \mathbb{R}^{C \times d}$ denotes the low-rank projection matrix.

Next, we encode the distance matrix $\mathbf{D}$ and inject it into the attention mechanism as a high-order structural bias term. In our approach, the attention mechanism employs three distinct matrices $(\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V) \in \mathbb{R}^{d \times d}$ to project $\mathbf{Z}$ and $\mathbf{Z}_p$. The process is detailed as follows:

$$\mathbf{Z}_G = \text{softmax}(\frac{\mathbf{Z}\mathbf{W}_Q(\mathbf{Z}_p\mathbf{W}_K)^T}{\sqrt{d}} + f_b(\mathbf{D}))\mathbf{Z}_p\mathbf{W}_V + \mathbf{Z}, \tag{11}$$

where $f_b(\cdot)$ is implemented as a linear transformation layer that maps shortest-path distances to attention biases. This attention mechanism integrates structural information directly into the global representation $\mathbf{Z}_G \in \mathbb{R}^{N \times d}$. It captures global context through its fully-connected nature, while simultaneously encoding high-order structural relationships between nodes using the low-rank projection $\mathbf{Z}_p$ and biases based on the distance matrix $\mathbf{D}$. Multi-head attention and residual connections with $\mathbf{Z}$ as the initial input are also employed to enhance model capacity.

To enable the downstream classification task to jointly leverage information from both global and local perspectives, we concatenate the global representation $\mathbf{Z}_G$ with the local representation $\mathbf{Z}_L$ and then project the combined features using learnable weight matrix $\mathbf{W}_c \in \mathbb{R}^{2d \times c}$,

$$\mathbf{Z}_f = (\mathbf{Z}_g \parallel \mathbf{Z}_l)\mathbf{W}_c, \tag{12}$$

where $\mathbf{Z}_f \in \mathbb{R}^{N \times c}$ denotes the final representation used for classification, and c is the number of classes.

Algorithm 1: Training Framework of CoGFormer.

Require: Multi-view data $\mathcal{X} = \{\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(V)}\}$, semi-supervised labels $\mathbf{Y} \in \mathbb{R}^{|\Phi| \times c}$.
Ensure: Final representation $\mathbf{Z}_f$.
1: Initialize weights $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ of the attention and learnable weights $\{\alpha^{(v)}\}_{v=1}^{V}$;
2: Initialize weights $\{\mathbf{W}_{gcn}^{(l)}\}_{l=1}^{L}$ of the graph convolutional network;
3: Construct adjacency matrices $\{\mathbf{A}^{(v)}\}_{v=1}^{V}$ via kNN;
4: Construct distance matrix $\mathbf{D}$ with (9);
5: **while** not convergent **do**
6: Compute shared representation $\mathbf{Z}$ and optimize $\{\mathbf{W}_e^{(l)}, \mathbf{W}_d^{(v,l)}\}_{l=1}^{L}$ by back propagation with (3);
7: Calculate local representation $\mathbf{Z}_L$ with (7);
8: Compute global representation $\mathbf{Z}_G$ with (11);
9: Obtain final representation $\mathbf{Z}_f$ with (12);
10: Update parameters by back propagation with (13);
11: **end while**
12: **return** Final representation $\mathbf{Z}_f$.

Training Details

To obtain the final predictive representation for the semi-supervised classification task, our method employs a loss function defined by the cross-entropy error:

$$\mathcal{L}_{ce} = -\sum_{i \in \Phi} \sum_{j=1}^{c} \mathbf{Y}_{ij} \ln \tilde{\mathbf{Z}}_{ij}, \tag{13}$$

where $\tilde{\mathbf{Z}} = \mathrm{softmax}(\mathbf{Z}_f)$, Φ denotes the set of labeled samples, and $\mathbf{Y} \in \mathbb{R}^{|\Phi| \times c}$ represents the one-hot labels of the labeled samples.

Our training strategy consists of three main steps: optimizing the parameters of the shared representation learning network, optimizing the parameters of the attention mechanism, and optimizing the parameters of the GCN augmented with the $\rho(\cdot)$ function. Each module performs one forward pass and one backward pass per iteration, updating its network parameters based on the loss functions $\mathcal{L}_{re}$ and $\mathcal{L}_{ce}$. Algorithm 1 presents the complete procedure of our method.

Experienment

Experiment Settings

Datasets We evaluate our method against seven compared methods on six widely used multi-view datasets. Table 2 presents the detailed information of these datasets used for testing.

Compared Methods We compare the proposed algorithm with seven compared baseline models, including CO-GCN (Li, Li, and Wang 2020), TMC (Han et al. 2021), LGCN-FF (Chen et al. 2023), IMvGCN (Wu et al. 2023b), ECML (Xu et al. 2024), GEGCN (Lu et al. 2024b), and ECMGD (Lu et al. 2024a).

Datasets	# Samples	# Views	# Classes	Data types
100leaves	1,600	3	100	Object images
ALOI	1079	4	10	Object images
Animals	10,158	2	50	Animal images
Caltech20	2,386	6	20	Object images
MNIST	10,000	3	10	Digit images
NoisyMNIST	15,000	2	10	Digit images

Table 2: A brief description of all test multi-view datasets.

Parameter Setting All baseline methods are tuned with optimal hyperparameters on each dataset to suit their respective tasks. Our method employs the Adam optimizer to update all learnable parameters. The learning rate lr is set to 0.001, and the weight decay is configured as 0.005. Both the shared encoder and multiple view-specific decoders within the framework are each composed of one layer. Two graph convolutional layers are utilized, with the hidden layer dimension d of the model fixed as 256. The sharpness coefficient δ of the shrinkage function is specified as 10, and the saturation threshold τ is set to 30. The initial adjacency matrix is constructed using k-nearest neighbors (kNN), where the hyperparameter k ranges from 15 to 30. The number of communities C generated via METIS partitioning ranges from 16 to 512. The dropout rate is fixed at 0.5 for all datasets.

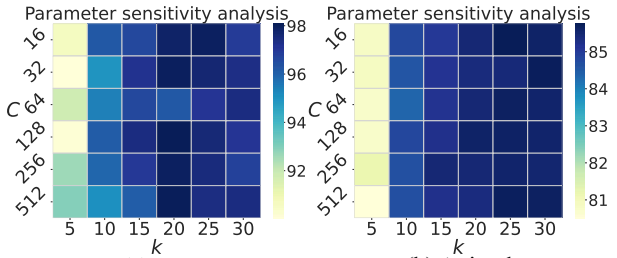
Experimental Results

Performance Evaluation We utilize 10% of the labeled samples for supervision and two classification evaluation metrics, including Accuracy (ACC) and F1-score are employed to assess the performance of these methods. Table 3 summarizes the performance of all methods across six datasets, leading to the following observations: 1) Through the joint learning of a GCN and an attention mechanism, CoGFormer achieves superior performance on all six datasets. This demonstrates the algorithm's superiority in semi-supervised classification and highlights the promise of graph transformers in multi-view learning; 2) CoGFormer exhibits superior classification performance on both the 100leaves and NoisyMNIST datasets, highlighting its capability to effectively capture inter-view associations across small-scale and large-scale datasets; 3) Compared with algorithms that utilize graph structure (such as Co-GCN, LGCN-FF, IMvGCN, and GEGCN), the proposed algorithm achieves the best performance, indicating that structure-consensus-enabled CoGFormer more effectively leverages structural information.

Parameters Analysis In this subsection, we investigate the impact of the hyperparameters k and C used in the proposed method. As shown in Figure 2, the model performance progressively improves and eventually stabilizes as the value of k increases. This indicates that the model can effectively leverage complex structures to generate consistent information, while maintaining robustness even in denser and noisier structural environments. Furthermore, results on both datasets demonstrate that variations in hyperparameter C do not lead to performance fluctuations. This suggests that the

Method	Metric	100leaves	ALOI	Animals	Caltech20	MNIST	NoisyMNIST
Co-GCN	ACC	46.70 (1.84)	84.22 (0.95)	77.64 (0.39)	80.74 (2.36)	92.10 (0.50)	90.10 (0.50)
	F1	42.39 (2.34)	84.56 (0.33)	71.03 (0.53)	52.42 (0.07)	91.94 (0.24)	89.73 (0.32)
TMC	ACC	60.86 (0.29)	86.79 (0.41)	83.62 (0.01)	66.58 (0.02)	89.92 (0.02)	87.21 (0.02)
	F1	59.14 (0.27)	87.20 (0.42)	77.25 (0.01)	37.39 (0.03)	89.71 (0.02)	86.71 (0.03)
LGCN-FF	ACC	88.53 (0.39)	95.51 (0.46)	81.25 (0.25)	74.39 (1.05)	88.68 (0.57)	89.74 (0.23)
	F1	88.29 (0.27)	95.52 (0.47)	70.94 (0.97)	38.81 (2.19)	85.56 (0.49)	89.57 (0.24)
IMvGCN	ACC	74.04 (2.18)	79.51 (6.55)	84.63 (0.01)	76.11 (0.14)	86.65 (0.68)	88.60 (0.10)
	F1	71.74 (2.13)	79.76 (6.99)	79.62 (0.11)	52.82 (0.23)	86.26 (0.78)	88.20 (0.10)
ECML	ACC	72.39 (0.39)	91.95 (0.52)	81.85 (0.02)	85.68 (0.16)	81.69 (0.41)	85.14 (0.01)
	F1	69.23 (0.49)	92.25 (0.45)	75.78 (0.02)	68.30 (1.06)	81.14 (0.51)	84.51 (0.03)
GEGCN	ACC	90.39 (0.39)	91.12 (0.21)	80.54 (0.04)	85.19 (0.38)	93.61 (0.03)	95.44 (0.05)
	F1	90.04 (0.46)	91.14 (0.15)	71.98 (0.06)	65.03 (0.94)	93.54 (0.03)	95.36 (0.06)
ECMGD	ACC	84.50 (0.29)	93.40 (0.72)	82.20 (0.13)	87.98 (0.19)	88.18 (0.04)	89.11 (0.02)
	F1	83.95 (0.29)	93.40 (0.74)	76.10 (0.20)	72.62 (0.68)	87.97 (0.03)	88.88 (0.02)
Ours	ACC	**96.93 (0.14)**	**98.04 (0.10)**	**85.76 (0.04)**	**90.24 (0.21)**	**93.78 (0.05)**	**97.06 (0.05)**
	F1	**96.91 (0.14)**	**98.05 (0.10)**	**80.50 (0.04)**	**77.58 (0.36)**	**93.72 (0.04)**	**97.01 (0.05)**

Table 3: Classification accuracy and F1-score (mean% and standard deviation%) of all methods. The best results are bolded and the sub-optimal results are underlined.



(a) ALOI (b) Animals

Figure 2: The parameter sensitivity analysis of hyperparameter k and C in CoGFormer on ALOI and Animals datasets.

Datasets\Methods	M3	M2	M1	CoGFromer
100leaves	94.96	95.21	95.71	**96.93**
ALOI	92.83	93.24	96.80	**98.04**
Animals	83.44	83.12	85.23	**85.76**
Caltech20	85.37	85.75	89.68	**90.24**
MNIST	93.47	93.25	93.61	**93.78**
NoisyMNIST	96.50	96.46	96.54	**97.06**

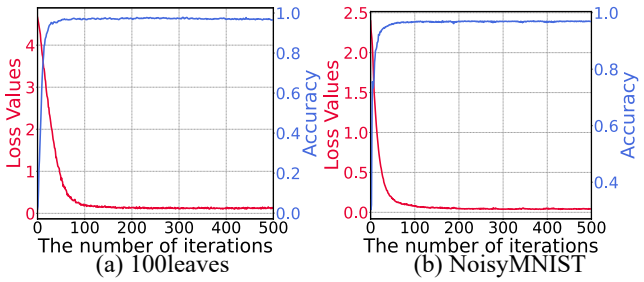
Table 4: Ablation study of the proposed method.

model exhibits robustness to hyperparameter C, while also confirming that the attention mechanism can maintain strong performance without increasing computational complexity.

Ablation Study To systematically investigate the impact of the structural consensus module on CoGFormer, we design three simple variants of its core components and conduct comparative experiments:

- M1: Replaces the structure-guided attention mechanism with a variant of the standard attention mechanism;
- M2: Employs an average-weighted graph convolution as a variant instead of the denoising structural consensus graph convolution;
- M3: Integrates a hybrid variant that combines the standard attention mechanism with average-weighted graph convolution.

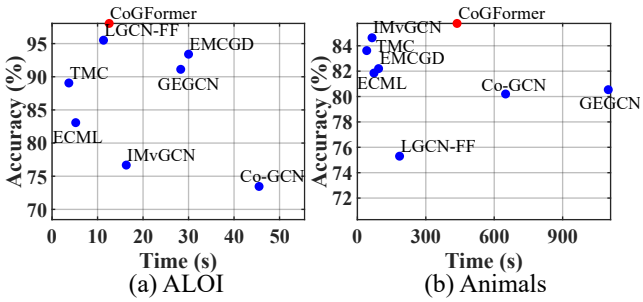
The experimental results, as shown in Table 4, demonstrate that the proposed approach achieves the best performance, which verifies the effectiveness of each module design and indicates that the model can leverage structural consensus information to achieve performance improvement.



(a) 100leaves (b) NoisyMNIST

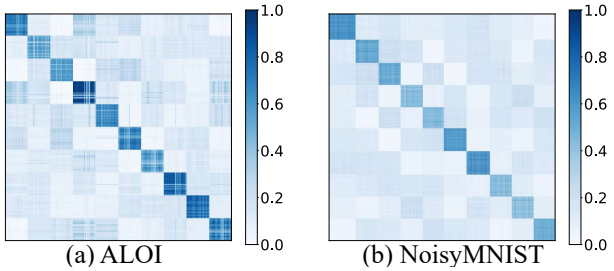
Figure 3: Curves of loss values (red), accuracy (blue) in CoGFormer on 100leaves and NoisyMNIST datasets.

Convergence and Runtime Analysis To validate the convergence of the proposed model, we analyze the trends of evaluation metrics and loss values across different test datasets. As illustrated in Figure 3, the loss values exhibit a



(a) ALOI (b) Animals

Figure 4: Running time (seconds) of the algorithms for 500 epochs on the ALOI and Animals datasets.



(a) ALOI (b) NoisyMNIST

Figure 5: Visualization of $\mathbf{Z}_f\mathbf{Z}_f^T$ on ALOI and NoisyMNIST.

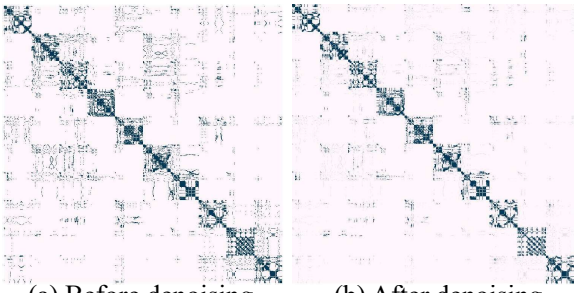
gradual decline and stabilization with increasing iterations, exhibiting a corresponding rapid increase in accuracy. Notably, the model demonstrates convergence within approximately 150 epochs.

In terms of model efficiency, since forward propagation is primarily governed by Equations (7) and (11), despite optimizations for attention computation efficiency, the model's time complexity remains $\mathcal{O}(N^2)$, as shown in Figure 4. Although computational overhead exists, this is compensated by the improvement in classification performance.

Visualization Analysis In this subsection, a similarity matrix is computed based on the learned representations, as shown in Figure 5. In the matrix, blue indicates higher values, while white indicates lower values. It can be observed from the figure that prominent blue block structures appear along the diagonal, indicating that instances within the same category exhibit high similarity, whereas instances from different categories show lower similarity.

Figure 6 presents the visualization results of denoising by the function $\rho(\cdot)$. In the right panel, it is evident that the structure selectively weakens minor connections between samples from different categories. This observation suggests that the model is capable of adaptively performing fine-grained manipulation of the graph structure.

Structural Consensus Analysis To assess structural consensus, we compare CoGFormer with its ablated variant (without structural consensus, CoGFormer_nsc) and other benchmarks that use graph structures, under varying structural complexity (Figure 7). Two key findings emerge: 1) As structural complexity (k) increases, CoGFormer's accuracy



(a) Before denoising (b) After denoising

Figure 6: Comparison before and after denoising with the function $\rho(\cdot)$.

rises and stabilizes, while others show minimal gains or even degradation, confirming that structural consensus effectively suppresses noise in complex environments; 2) Even without this mechanism, CoGFormer_nsc outperforms most benchmarks, underscoring the inherent advantage of graph transformers in capturing global receptive fields.

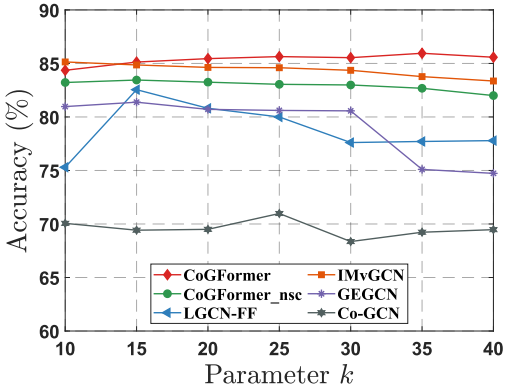


Figure 7: Accuracy trends of different algorithms on the Animals dataset as the hyperparameter k value increases.

Conclusion

In this paper, we propose a multi-view learning model named CoGFormer, which integrates a denoising structural consensus GCN and a structure-guided attention mechanism. By combining local and global information and designing structural consensus, it provides a unified framework that effectively utilizes structural information in complex multi-view topological environments. A series of experiments validate the performance of the proposed approach. We believe that CoGFormer would offer new insights into the integration of graph transformers in multi-view learning and lay a solid foundation for developing more robust and generalizable multi-view representation models.

Acknowledgments

This work is in part supported by the National Natural Science Foundation of China under Grants U25A20527 and 62276065, and the Fujian Provincial Natural Science Foundation of China under Grant 2024J01510026.

References

Chen, D.; O'Bray, L.; and Borgwardt, K. M. 2022. Structure-aware transformer for graph representation learning. In *Proceedings of the Thirty-Ninth International Conference on Machine Learning*, 3469–3489.

Chen, Z.; Fu, L.; Yao, J.; Guo, W.; Plant, C.; and Wang, S. 2023. Learnable graph convolutional network and feature fusion for multi-view learning. *Information Fusion*, 95: 109–119.

Cheng, J.; Wang, Q.; Tao, Z.; Xie, D.; and Gao, Q. 2020. Multi-view attribute graph convolution networks for clustering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2973–2979.

Cheng, Z.; Zhong, T.; Zhang, K.; Walker, J.; and Zhou, F. 2022. Learning contrastive multi-view graphs for recommendation (student abstract). In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, 12927–12928.

Gui, A.; Ye, J.; and Xiao, H. 2024. G-Adapter: Towards structure-aware parameter-efficient transfer learning for graph transformer networks. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, 12226–12234.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2021. Trusted multi-view classification. In *Proceedings of the Ninth International Conference on Learning Representations*, 1–16.

Karypis, G.; and Kumar, V. 1998. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing*, 20(1): 359–392.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the Fifth International Conference on Learning Representations*, 1–14.

Kong, X.; Liu, J.; Li, H.; Zhang, C.; Du, J.; Guo, D.; and Shen, G. 2025. Graph anomaly detection via diffusion enhanced multi-view contrastive learning. *Knowledge-Based Systems*, 311: 113093.

Li, Q.; Ni, H.; and Wang, Y. 2024. NHGMI: Heterogeneous graph multi-view infomax with node-wise contrasting samples selection. *Knowledge-Based Systems*, 289: 111520.

Li, S.; Li, W.; and Wang, W. 2020. Co-gcn for multi-view semi-supervised learning. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 4691–4698.

Li, Y.; Bao, H.; Ge, Z.; Yang, J.; Sun, J.; and Li, Z. 2023. BEVStereo: Enhancing depth estimation in multi-view 3D object detection with temporal stereo. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, 1486–1494.

Liang, W.; Liu, X.; Zhou, S.; Liu, J.; Wang, S.; and Zhu, E. 2022. Robust graph-based multi-view clustering. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, 7462–7469.

Liu, C.; Zhan, Y.; Ma, X.; Ding, L.; Tao, D.; Wu, J.; and Hu, W. 2023. Gapformer: Graph transformer with graph pooling for node classification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2196–2205.

Lou, W.; Li, G.; Wan, X.; and Li, H. 2024. Cell graph transformer for nuclei classification. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, 3873–3881.

Lu, J.; Wu, Z.; Chen, Z.; Cai, Z.; and Wang, S. 2024a. Towards multi-view consistent graph diffusion. In *Proceedings of the Thirty-Second ACM International Conference on Multimedia*, 186–195.

Lu, J.; Wu, Z.; Zhong, L.; Chen, Z.; Zhao, H.; and Wang, S. 2024b. Generative essential graph convolutional network for multi-view semi-supervised classification. *IEEE Transactions on Multimedia*, 26: 7987–7999.

Luo, Y.; Li, H.; Shi, L.; and Wu, X. 2024. Enhancing graph transformers with hierarchical distance structural encoding. In *Proceedings of the Thirty-Eighth Conference on Neural Information Processing Systems*, 1–33.

Peng, T.; Liang, Y.; Wu, W.; Ren, J.; Pengrui, Z.; and Pu, Y. 2023. CLGT: A graph transformer for student performance prediction in collaborative learning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, 15947–15954.

Pu, J.; Cui, C.; Chen, X.; Ren, Y.; Pu, X.; Hao, Z.; Yu, P. S.; and He, L. 2024. Adaptive feature imputation with latent graph for deep incomplete multi-view clustering. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, 14633–14641.

Tang, C.; Liu, X.; Zhu, X.; Zhu, E.; Luo, Z.; Wang, L.; and Gao, W. 2020. CGD: Multi-view clustering via cross-view graph diffusion. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 5924–5931.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the Thirty-First Conference on Neural Information Processing Systems*, 5998–6008.

Wang, Q.; Xu, H.; Zhang, Z.; Feng, W.; and Gao, Q. 2025. Deep multi-modal graph clustering via graph transformer network. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, 7835–7843.

Wang, S.; Huang, S.; Wu, Z.; Liu, R.; Chen, Y.; and Zhang, D. 2024. Heterogeneous graph convolutional network for multi-view semi-supervised classification. *Neural Networks*, 178: 106438.

Wen, Z.; Ling, Y.; Ren, Y.; Wu, T.; Chen, J.; Pu, X.; Hao, Z.; and He, L. 2024. Homophily-Related: Adaptive hybrid graph filter for multi-view graph clustering. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, 15841–15849.

Wu, Q.; Yang, C.; Zhao, W.; He, Y.; Wipf, D.; and Yan, J. 2023a. DIFFormer: Scalable (graph) transformers induced by energy constrained diffusion. In *Proceedings of the Eleventh International Conference on Learning Representations*, 1–26.

Wu, Z.; Lin, X.; Lin, Z.; Chen, Z.; Bai, Y.; and Wang, S. 2023b. Interpretable graph convolutional network for multi-view semi-supervised learning. *IEEE Transactions on Multimedia*, 25: 8593–8606.

Xiong, F.; Zhang, T.; Pan, S.; Luo, G.; and Wang, L. 2025. Robust graph based social recommendation through contrastive multi-view learning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, 12890–12898.

Xu, C.; Si, J.; Guan, Z.; Zhao, W.; Wu, Y.; and Gao, X. 2024. Reliable conflictive multi-view learning. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, 16129–16137.

Yang, D.; Huang, S.; Xu, Z.; Li, Z.; Wang, S.; Li, M.; Wang, Y.; Liu, Y.; Yang, K.; Chen, Z.; Wang, Y.; Liu, J.; Zhang, P.; Zhai, P.; and Zhang, L. 2023. AIDE: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20402–20413.

Yu, P.; Tan, Z.; Lu, G.; and Bao, B. 2023. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the Thirty-First ACM International Conference on Multimedia*, 6576–6585.

Zhang, X.; Cai, F.; Zheng, J.; Pan, Z.; Chen, W.; Chen, H.; and Chen, C. 2025. Triangle Matters! TopDyG: Topology-aware transformer for link prediction on dynamic graphs. In *Proceedings of the ACM on Web Conference*, 3607–3617.

Zhang, Z.; Liu, Q.; Hu, Q.; and Lee, C. 2022. Hierarchical graph transformer with adaptive node sampling. In *Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems*, 1–13.

Zhao, L.; Wang, Z.; Wang, X.; Chen, Z.; and Xu, B. 2025. Incomplete and unpaired multi-view graph clustering with cross-view feature fusion. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, 22786–22794.

Zheng, Y.; Gindra, R. H.; Green, E. J.; Burks, E. J.; Betke, M.; Beane, J. E.; and Kolachalama, V. B. 2022. A graph-transformer for whole slide image classification. *IEEE Transactions on Medical Imaging*, 41: 3003–3015.

Zhou, P.; and Du, L. 2023. Learnable graph filter for multi-view clustering. In *Proceedings of the Thirty-First ACM International Conference on Multimedia*, 3089–3098.

Zhou, W.; Haq, A. U.; Qiu, L.; and Akbar, J. 2024. Multi-view social recommendation via matrix factorization with sub-linear convergence rate. *Expert Systems with Applications*, 237(Part C): 121687.