U2B: Scale-unbiased Representation Converter for Graph Classification with Imbalanced and Balanced Scale Distributions

Anonymous submission

Abstract

Graph classification is a critical task in analyzing graph data, with applications across various domains. While graph neural networks (GNNs) have achieved remarkable results, their ability to generalize across graphs of varying scales remains a challenge. Conventional models often perform well on large-scale graphs but struggle with distributions that are skewed towards small scales. Conversely, models tailored to address scale imbalances frequently prioritize small-scale graphs, leading to diminished performance in more balanced scenarios. To overcome these limitations, we introduce a **U**nbalanced-**B**alanced Representation Converter (U2B), which exhibits no explicit bias toward graph scales. U2B employs a two-step workflow: a distillation phase to extract base features from both node-level and graph-level representations, followed by a refinement phase to generate biased representations for improved balance. In the distillation phase, a static constraint guides node-level adjustments, improving the representation of nodes in small graphs. Simultaneously, a dynamic constraint in the graph-level process mitigates biases toward features from large graphs. To ensure harmony between the representations, a consistency alignment loss is introduced, aligning node-level and graph-level features to create more cohesive and balanced graph representations. Extensive experiments with **16** baselines across **8** datasets demonstrate that U2B achieves SOTA performance, boasting improvements of up to **22.19%**. Additionally, we establish its strong compatibility with a range of other models. All associated code is provided in **Supplement**.

1 Introduction

Graph classification is essential in graph data analysis, focusing on predicting a graph's category based on its structural topology and node attributes (Wu et al. 2020). In recent years, Graph Neural Networks (GNNs) have emerged as a dominant approach for this task due to their robust message-passing mechanisms, which effectively capture and represent hierarchical graph structures (Wei et al. 2023; Xie et al. 2022).

In real-world applications, graph data often exhibit complex variations in scale. As shown in Figure 1(a), three commonly used graph classification datasets highlight these differences. For instance, datasets like PROTEINS and PTC-MR follow a power-law distribution, characterized by a few large "head" graphs and numerous small "tail" graphs, forming a long-tail pattern. In contrast, the NCI1 dataset displays a normal distribution, lacking the long-tail characteristics.

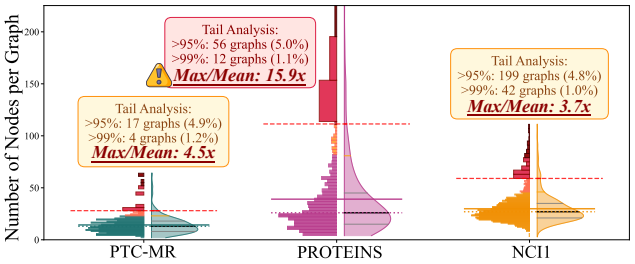


Figure 1: Graph-scale distribution of different datasets.

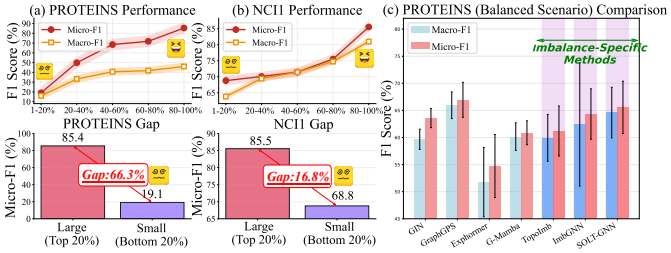


Figure 2: Graph classification performance of GIN across different graph scales on PROTEINS (a) and NCI1 (b) datasets as examples. (c) shows performance comparison between scale imbalance-specific methods and conventional models.

The diverse scale distributions in graph datasets pose significant challenges to the generalization capabilities of existing graph classification methods. Conventional approaches often prioritize large-scale graphs while neglecting small-sacle ones, resulting in poor performance on datasets with long-tail characteristics. As illustrated in Figure 2(a) and (b), the predictive performance of GIN varies significantly across different graph scales in the PROTEINS and NCI1 datasets, with further details provided in Table 4. On the other hand, specific models designed for small-scale graphs frequently sacrifice performance on large-sacle graphs, leading to suboptimal results on datasets without long-tail distributions. For example, Figure 2(c) shows that such methods can sometimes underperform compared to standard GIN. Consequently, the development of a robust approach that maintains strong performance across both balanced and imbalanced graph scale distributions remains an urgent challenge.

To address these challenges, we propose the **U**nbalanced-**B**alanced Representation Converter (U2B). The key idea behind U2B is to distill scale-invariant base features, which are then used to reconstruct biased node-level and graph-level representations. This process creates a scale-unbiased feature space, ultimately improving performance across diverse graph classification scenarios.

U2B features a structured approach centered on extracting and reconstructing key features. The method begins by establishing learnable compressed tokens, which are used in a distillation process to adaptively extract fundamental features from input samples. These features are then refined to reconstruct representations that effectively capture the underlying graph structure. At the node level, we introduce a static balancing constraint during the distillation stage to ensure balanced activation of the compressed tokens. This prevents overemphasis on nodes from large graphs and enhances the model's ability to recognize structural patterns in small graphs. The resulting node representations are pooled to generate graph representations, which form the input for the next stage. At the graph level, we employ dynamic balancing constraints to align feature distributions across graphs of varying scales. This fosters the creation of unbiased representations that improve prediction accuracy across both balanced and imbalanced datasets. To further enhance the quality of the representations, we incorporate a consistency alignment loss, promoting compactness and semantic coherence. The primary contributions of this work are as follows:

❶ To the best of our knowledge, this is the first study to systematically address graph classification across comprehensive graph scale distributions, effectively handling both balanced and imbalanced settings.

❷ We propose a novel approach, U2B, designed to generate scale-unbiased graph representations through a two-stage framework consisting of distillation and refinement. This approach incorporates two specific constraint terms to regulate the distillation process, ensuring the extraction of representative fundamental features.

❸ We demonstrate the effectiveness and broad applicability of U2B through comparative experiments on **8** datasets and **16** baselines, with accuracy gains of up to **22.19%**.

2 Related Work

Graph Representation Learning Graphs are fundamental structures for modeling complex relationships. Graph representation learning encodes graph data into vector forms and has been proven effective in supporting various downstream tasks, such as graph classification (Hamilton 2020). Early approaches mainly relied on spectral methods, such as Laplacian operators (Noutahi et al. 2019) and random walks (Verdier et al. 2021). The success of deep learning in Euclidean domains has led to the emergence of graph neural networks (GNNs). GNNs leverage a message-passing framework that combines node features with graph topology, enabling nodes to exchange information while preserving both structural and content-related information. Inspired by the success of Transformer, recent research has extended these techniques to graphs, enabling more effective capture of global dependencies and structural characteristics (Yun et al. 2019; Cai and Lam 2020). However, these methods have demonstrated suboptimal performance in graph classification scenarios with topology-imbalanced graphs.

Imbalanced Learning for Node Classification Recent research has shifted toward investigating the long-tail problem within graph structures. For node-level imbalance learning tasks, methods such as DR-GCN (Shi et al. 2020) alleviate multi-class imbalance through class-conditional adversarial training, GraphSMOTE (Zhao, Zhang, and Wang 2021) performs over-sampling of minority nodes to generate new representations, and ImGAGN (Qu et al. 2021) leverages generative adversarial networks to simulate the attributes and structures of tail nodes. However, these methods require fine-grained node labels, which is fundamentally different from graph-level imbalance learning tasks, where the focus is on predicting a single label for the entire graph.

Imbalanced Learning for Graph Classification Graph imbalanced classification tasks can be further subdivided into class imbalance and topological imbalance. Methods designed for class imbalance emphasize learning from tail classes. For instance, G^2GNN (Wang et al. 2022) and ImbGNN (Xu et al. 2024) achieve data augmentation by fusing head and tail graphs. In this work, we focus on the fine-grained phenomenon of scale imbalance, which is prevalent in graph datasets. TopoImb (Zhao et al. 2022) mitigates structural imbalance by integrating topology-aware modeling and instance weighting, while SOLT-GNN (Liu et al. 2022) balances learning across graphs of different sizes by identifying co-occurring subgraph patterns. Despite these advances, such specialized models often sacrifice performance on large-scale graphs, leading to suboptimal results on datasets with non-long-tail distributions.

3 Problem Formulation

Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{A}, X\}$ be a graph, where $\mathcal{V}$ is the node set with $|\mathcal{V}|$ nodes, $\mathcal{E}$ is the edge set, $\mathcal{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is the adjacency matrix, $X \in \mathbb{R}^{|\mathcal{V}|\times d}$ is the node feature matrix with d-dimension. Given a graph set $\mathbb{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_N\}$ with N graphs with their corresponding label set $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_N\}$, the goal of graph-level classification task is to learn a mapping function $\mathcal{F} : \mathcal{G} \rightarrow \mathbb{R}^f$ to map any graph to a low-dimensional vector $h \in \mathbb{R}^f$. This representation is subsequently processed by a classifier to generate the predicted label distribution, resulting in the final output prediction for the graph sample.

4 Method

Overall Architecture of U2B

As shown in Figure 3, U2B integrates both node-level and graph-level unbalanced-to-balanced transformers to reshape the generated representations, ultimately producing scale-unbiased representations. The transformer extracts fundamental features from unbalanced data and employs specific strategies to constrain this process in a balanced manner, effectively mitigating learning biases. Specifically, for a batch of input graphs, the adjacency matrices and the feature matrix

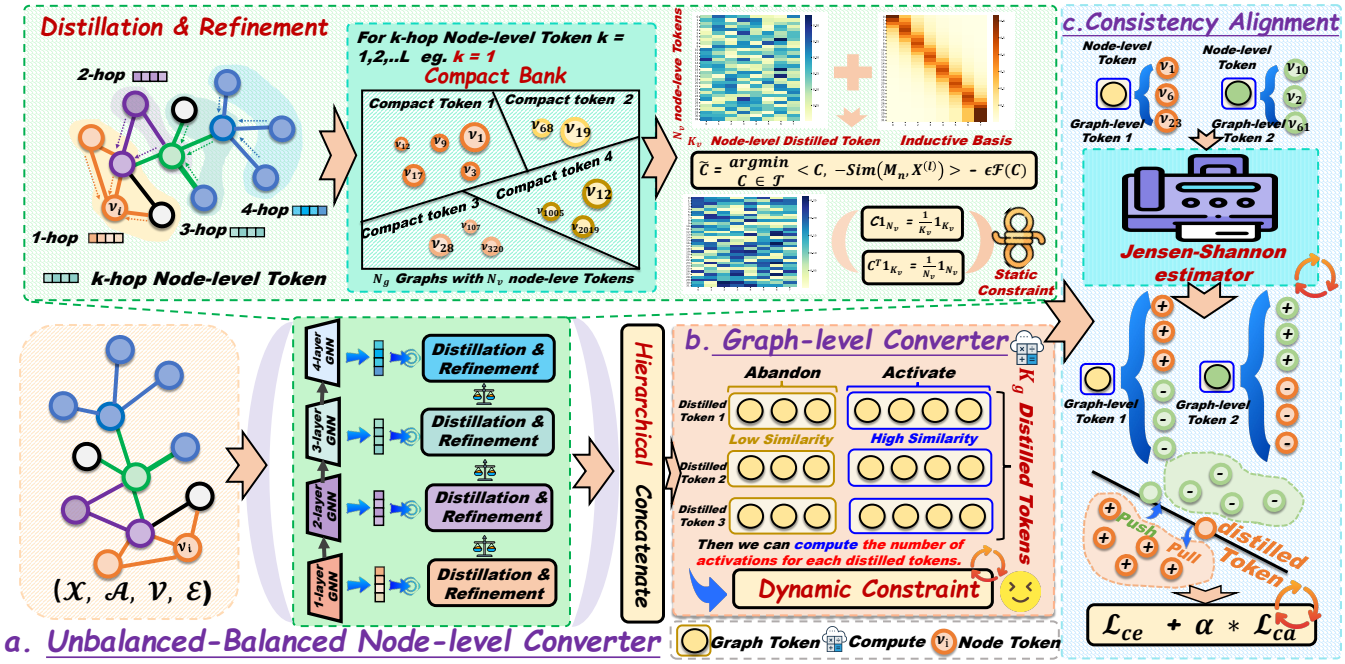


Figure 3: Overview of U2B, including (a) Node-level Converter, (b) Graph-level Converter, (c) Consistency Alignment Loss.

are concatenated as $\mathbf{A} \in \mathbb{R}^{N_v \times N_v}$ and $\mathbf{X} \in \mathbb{R}^{N_v \times d}$, where N_v denotes the total number of nodes in the batch and d is the feature dimension. We stack GNN encoders (e.g., GIN (Xu et al. 2018)) to capture complex feature interactions among nodes, thereby generating node-level representations. These representations from each GNN layer are then input to the node-level unbalanced-balanced converter. The output of this converter is subsequently passed through a READOUT function to aggregate node features into a graph-level representation $\mathbf{H}$. The graph-level converter refines this representation, and then outputs scale-unbiased graph representation $\widetilde{\mathbf{H}}$. Finally, a classifier is employed to predict the labels. The final loss integrates the classification loss $\mathcal{L}_{ce}$ and the proposed consistency loss $\mathcal{L}_{ca}$.

Unbalanced-Balanced Node-Level Representation

GNNs for Representation. GNNs are a class of deep learning models designed to operate on graph-structured data, with notable examples including GCN (Kipf and Welling 2016) and GIN (Xu et al. 2018). A key operation in GNNs is neighborhood aggregation, where each node recursively gathers and transforms information from its neighbors to construct its own representation. Given the input in the l-th GNN layer $\mathbf{X}^{(l-1)}$ with $\mathbf{X}^{(0)} = \mathbf{X}$, it can be formally expressed as follows:

$$\mathbf{X}_v^{(l)} = \mathbf{X}_v^{(l-1)} + \mathrm{AGG}\left(\left\{\left(\mathbf{X}_v^{(l-1)}, \mathbf{X}_u^{(l-1)}\right) : \mathcal{U} \in \mathcal{N}(v)\right\}\right) \tag{1}$$

where $\mathbf{X}_v^{(l)}$ means the representation of node v at l-th layer. And $\mathcal{N}(v)$ denote the set of neighbors of node v in the graph. The function AGG $(\cdot)$ represents an aggregation operation such as a weighted average or attention mechanism.

Node-level Converter. In GNN message passing, nodes in small graphs often aggregate insufficient neighborhood information, resulting in under-informed embeddings. To mitigate this issue, the distillation stage in U2B compresses node representations into coarse-grained base features that capture shared patterns across graphs. In the refinement stage, these features are decoded into fine-grained, node-specific representations, effectively restoring representational balance.

Specifically, given the output of l-th GNN layer $\mathbf{X}^{(l)} \in \mathbb{R}^{N_v \times d}$, where N_v means the number of nodes in the graph and d is the number of channels, the converter introduces a compact bank of node-level distilled token representations, which is denoted as $\mathbf{M}_n \in \mathbb{R}^{K_v \times d}$. These tokens distill base features from input. And hyperparameter K is the number of compact tokens and set to $K_v \ll N_v$.

❶ *Distillation Process.* We employ the self-attention mechanism to compute the similarity between the distilled token representation and node representations, enabling the model to adaptively perceive their shared components as follows,

$$\mathrm{Sim}(\mathbf{M}_n, \mathbf{X}^{(l)}) = \mathrm{Softmax}\left(\mathbf{M}_n (W_q^{(l)} \mathbf{X}^{(l)})^\top / \sqrt{d} + \psi\right) \tag{2}$$

where the generated distillation coefficient matrix is denoted as $\mathbf{Att}_n \in \mathbb{R}^{K_v \times N_v}$, and $W_q^{(l)}$ is the learnable parameter. $\psi = W^p (\mathbf{E})^\top \in \mathbb{R}^{K_v \times N_v}$, where $W^p \in \mathbb{R}^{K_v \times 3}$ is the learnable parameter and $\mathbf{E} \in \mathbb{R}^{N_v \times 3}$ encodes the degree of each node, the total number of nodes in the graph to which it belongs, and the average degree of that graph, which serves as an inductive bias to guide the subsequent distillation process.

The distillation process relies on fine-grained interactions between nodes and coarse-grained entities. However, it can face the challenge of "coarse collapse," where most nodes

are assigned to only a few clusters, leaving the remaining clusters sparsely populated or empty (Caron et al. 2018; Zeng et al. 2023). This imbalance obstructs the effective extraction of base patterns. Inspired by prior work (Caron et al. 2020), we incorporate a Sinkhorn-Knopp-based static constraint to ensure balanced access to each distilled token during the distillation process, preventing collapse and improving pattern extraction. Specifically, the Sinkhorn-Knopp strategy begins by establishing a constraint matrix, $\mathbf{C} \in \mathbb{R}^{K_v \times N_v}$. This matrix serves as a guide for the distillation process, ensuring balanced interactions between coarse-grained tokens and fine-grained nodes. The objective function is then formulated and optimized as follows:

$$\tilde{\mathbf{C}} = \underset{\mathbf{C} \in \mathcal{T}}{\operatorname{argmin}} \left\langle \mathbf{C}, -\operatorname{Sim}\left(\mathbf{M}_n, \mathbf{X}^{(l)}\right) \right\rangle - \epsilon \mathcal{F}(\mathbf{C}) \quad (3)$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product between two matrices, $\mathcal{F}(\cdot)$ is the entropy function and $\mathcal{F}(\mathbf{C}) = -\sum_{\mathrm{i}=1}^{\mathrm{K}_v} \sum_{\mathrm{j}=1}^{\mathrm{N}_v} \mathbf{C}_{\mathrm{ij}} \log \mathbf{C}_{\mathrm{ij}}$, $\mathbf{C}_{\mathrm{ij}}$ denotes the element in the i-th row and j-th column of $\mathbf{C}$. ϵ is a hyperparameter that controls the smoothness of the learning. Following previous work (Asano, Rupprecht, and Vedaldi 2019), we enforce equal partitioning by constraining the matrix $\mathbf{C}$ within the transportation polytope. We propose to adapt their solution to mini-batch data by restricting the transportation polytope to each mini-batch:

$$\mathcal{T} = \left\{ \mathbf{C} \in \mathbb{R}_{+}^{K_v \times N_v} \;\middle|\; \mathbf{C}\mathbf{1}_{N_v} = \frac{1}{\mathrm{K}_v}\mathbf{1}_{K_v}\,, \mathbf{C}^{\mathrm{T}}\mathbf{1}_{K_v} = \frac{1}{\mathrm{N}_v}\mathbf{1}_{N_v} \right\} \quad (4)$$

where $\mathbf{1}_{K_v}$ denotes the vector of ones in dimension K_v. These constraints enforce that on average each distilled token is selected at least $\frac{N_v}{K_v}$ times. In fact, solving Equation 3 corresponds to an entropy-regularized optimal transport problem. The work (Caron et al. 2020) demonstrates that obtaining a continuous solution to this problem yields better performance. Therefore, the solution $\mathbf{C}^{\star}$ to the optimization problem of Equation 3 can be approximated by a soft assignment matrix that is normalized along both rows and columns.

$$\tilde{\mathbf{C}} = \operatorname{Diag}(\mathbf{u}) \exp\left(\operatorname{Sim}(\mathbf{M}_n, \mathbf{X}^{(l)})/\epsilon\right) \operatorname{Diag}(\mathbf{v}) \quad (5)$$

where $\mathbf{u}$ and $\mathbf{v}$ are renormalization vectors in $\mathbb{R}^{K_v}$ and $\mathbb{R}^{N_v}$ respectively, and $\epsilon > 0$ is the entropy regularization coefficient, which controls the smoothness of the transport matrix and it is a fixed value in our case.

❷ *Refinement Process.* This process entails mapping the extracted coarse-grained base representations back to fine-grained node representations during distillation, for which we continue to employ the attention mechanism.

Finally, the computation of the node-level converter, which includes the distillation and refinement processes, can be formulated as follows,

$$\widetilde{\mathbf{X}}^{(l)} = \operatorname{Sigmoid}\left(W_q^{(l)}\mathbf{X}^{(l)}\left(\mathbf{M}_n\right)^{\top}/\sqrt{d}\right)\tilde{\mathbf{C}}W_v^{(l)}\mathbf{X}^{(l)} \quad (6)$$

where $W_q^{(l)}$ and $W_v^{(l)}$ are learnable parameters. $\widetilde{\mathbf{X}}^{(l)} \in \mathbb{R}^{N_v \times d}$ is the calibrated output representation, which will serve as the input for the next layer $\mathbf{X}^{(l+1)}$. Finally, we stack L layers of GNN and concatenate the outputs of each layer to generate the final node representations:

$$\overline{\mathbf{X}} = \operatorname{CONCAT}\left(\left\{\widetilde{\mathbf{X}}^{(l)}\right\}_{l=1}^{L}\right) \in \mathbb{R}^{N_v \times (L \times d)} \quad (7)$$

Unbalanced-Balanced Graph-Level Representation

Graph-level Converter. After passing through L GNN layers, the output node representations $\overline{\mathbf{X}}$ are pooled using the READOUT function (Sun et al. 2019) to generate the graph-level representation $\mathbf{H} \in \mathbb{R}^{N_g \times d}$, where N_g represents the number of input graphs. Despite pooling, the representations of large-scale and small-scale graphs still differ in information content, which may lead to biased learning in the final model. To address this, we introduce a two-stage graph-level converter that distills essential features at the graph level to enhance the representations of small-sacle graphs. Analogous to previous steps, the process begins by initializing the graph-level distilled token representations $\mathbf{M}_g \in \mathbb{R}^{K_g \times d}$.

❶ *Distillation Process.* The graph-level distilled attention is defined as follows,

$$\operatorname{Sim}(\mathbf{M}_g, \mathbf{H}) = \operatorname{Softmax}\left(\operatorname{TopK}\left(\mathbf{M}_g\left(W_q^g\mathbf{H}\right)^{\top}/\sqrt{d} + \gamma\right)\right) \quad (8)$$

where W_q is the learnable parameter. And we retain the TopK largest attention coefficients, as the number of graphs is much smaller than the number of nodes, making this approach more effective for distilling the base features. The generated distillation coefficient matrix is denoted as $\mathbf{Att}_g \in \mathbb{R}^{K_g \times N_g}$. $\gamma \in \mathbb{R}^{K_g \times N_g}$ is a dynamic constraint term designed to prevent attention from being concentrated on only a few tokens.

Specifically, we first calculate the number of times each distillation token is activated. For k-th distilled token, u_k means the number of nonzero elements in the k-th row of the matrix $\mathbf{Att}_g$. To achieve balanced activation of the distillation tokens, the expected activation frequency for each distilled token is set to $\widetilde{\mu} = \frac{N_g \times \mathrm{TopK}}{K_g}$. γ captures the difference between $\widetilde{\mu}$ and μ, incorporating it as an intermediate term in the learning process. Since the discrete number of activations is non-differentiable, the update of γ is optimized subject to the following constraints as follows,

$$\gamma \leftarrow \gamma - \eta \nabla \mathcal{G}\left(\widetilde{\mu}, \mu\right), \quad (9)$$

$$\nabla \mathcal{G}\left(\widetilde{\mu}, \mu\right) = \begin{cases} \widetilde{\mu} - \mu, & \text{if } |\widetilde{\mu} - \mu| \leq \delta \\ \delta \operatorname{sign}\left(\widetilde{\mu} - \mu\right), & \text{otherwise} \end{cases} \quad (10)$$

where η is the learning rate, and δ is smoothing threshold. $\operatorname{sign}(\cdot)$ is a function of the numerical sign. When the activation count of a particular token exceeds the expected average, γ imposes an additional penalty on the subsequent attention coefficients, thus maintaining the activation of other tokens.

Unlike enforcing strict balance on the distillation coefficients in the node-level converter using the Sinkhorn algorithm, we adopt the aforementioned dynamic constraint strategy to adjust the attention weights in a more flexible manner, thereby capturing the representative base patterns in the graph structure more effectively. We conduct ablation studies on these two constraint mechanisms in **Supplement**

to evaluate their performance and applicability in different scenarios.

❷ *Refinement Process.* Similarly, during the distillation process, we map the extracted coarse-grained base representations back to fine-grained node representations, for which we continue to use the TopK attention mechanism.

Finally, the calculation of the graph-level converter can be written as,

$$\widetilde{\mathbf{H}} = \text{Sigmoid}\left(\text{TopK}\left(W_q^g \mathbf{H}\left(\mathbf{M}_g\right)^\top / \sqrt{d}\right)\right) \text{Sim}(\mathbf{M}_g, \mathbf{H}) W_v^g \mathbf{H} \tag{11}$$

Then, the generated representation $\widetilde{\mathbf{H}}$ is fed into a decoder to predict the label of the graphs.

Optimizing Loss with Consistency Alignment. To generate compact graph representations, we propose a consistency constraint optimization that aligns node-level representations with their corresponding graph-level representations while repelling the graph-level representations of non-corresponding graphs. This loss can be formulated as follows,

$$\begin{aligned}\mathcal{L}_{ca} = -\sum_{\mathcal{G}\in\mathbb{G}} \frac{1}{|\mathcal{G}|} \sum_{\mathcal{V}\in\mathcal{G}} \Big[\mathbb{E}\left[-\log\left(1+\mathrm{e}^{-\Gamma\left(\overline{\mathbf{X}}_{\mathcal{V}}, \widetilde{\mathbf{H}}_{\mathcal{G}}\right)}\right)\right] \\ +\mathbb{E}\left[\log\left(1+\mathrm{e}^{-\Gamma\left(\overline{\mathbf{X}}_{\mathcal{V}}, \widetilde{\mathbf{H}}_{\mathcal{G}'}\right)}\right)\right]\Big]\end{aligned} \tag{12}$$

where $\overline{\mathbf{X}}_{\mathcal{V}} \in \mathbb{R}^d$ represents the representation of node $\mathcal{V}$. $\widetilde{\mathbf{H}}_{\mathcal{G}}$ and $\widetilde{\mathbf{H}}_{\mathcal{G}'}$ denote the representation of the graph to which node $\mathcal{V}$ belongs and the representation of non-corresponding graphs, respectively. $\Gamma(\cdot)$ is the Jensen-Shannon estimator (Nowozin, Cseke, and Tomioka 2016) to express the affinity between two representations. The final loss function integrates the cross-entropy classification loss $\mathcal{L}_{ce}$ and the consistency alignment loss $\mathcal{L}_{ca}$:

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha * \mathcal{L}_{ca} \tag{13}$$

where α is the loss balance factor.

5 Experiment

Experiment Setup

► **Datasets.** We conduct graph classification experiments on 8 real-world datasets including 6 binary-class datasets and 2 multi-class datasets. These datasets are collected from various domains such as chemistry, protein biology, and social networks. We summarized the statistics in Table 1. More detailed information is provided in the **Supplement**.

Dataset	# Graphs	Avg. #Nodes	Avg. #Edges	# Features	# Classes	Avg. #Degree
PTC-MR	344	14.29	14.69	18	2	2.06
IMDB-BINARY	1000	19.77	96.53	65	2	9.77
PROTEINS	1113	39.06	72.82	3	2	3.73
D&D	1178	284.32	715.66	89	2	5.03
NCI1	4110	29.87	32.30	37	2	2.16
REDDIT-B	2000	429.63	497.75	-	2	2.32
COLLAB	5000	74.49	2457.22	-	3	65.97
IMDB-MULTI	1500	13.00	65.94	-	3	10.14

Table 1: Statistics of the used graph classification datasets.

► **Baselines.** Our experiments include six categories of baseline methods, comprising a total of **16** models. These methods include *(1) Graph kernel methods*: GK (Graphlet Kernel) (Shervashidze et al. 2009) and WL (Weisfeiler-Lehman subtree kernel) (Shervashidze et al. 2011). *(2) Classic graph neural networks*: GIN (Xu et al. 2018), GIN$^+$ (Luo, Shi, and Wu 2025), GraphSAGE, and GCN. *(3) Advanced graph neural networks*: DiffPool (Ying et al. 2018),CurGraph (Wang et al. 2021a), and Mixup (Wang et al. 2021b), and DGCNN (Zhang et al. 2018). *(4) Graph contrastive learning methods*: InfoGraph (Sun et al. 2019) and GraphCL (You et al. 2020). *Graph Transformer methods*: GraphGPS (Rampášek et al. 2022), Exphormer (Shirzad et al. 2023), and GraphMamba (Wang et al. 2024). *(5) Topology imbalance-oriented methods*: TopoImb (Zhao et al. 2022), ImbGNN (Xu et al. 2024), and SOLT-GNN (Liu et al. 2022). In this paper, we use a classic graph neural network GIN as the feature extractor for our main experiments. Further details of these baselines can be found in **Supplement**.

► **Setting.** We conduct evaluations under both ***scale-imbalanced*** and ***scale-balanced*** scenarios, following the protocols of the latest graph imbalance learning benchmark, IGL-Bench (Qin et al. 2024). For each dataset, the graphs with the largest 20% scale are defined as *large-scale*, while the remaining 80% are defined as *small-scale*. The statistics of the processed datasets are summarized in **Supplement**. We split each dataset into 10% for training, 10% for validation, and 80% for testing. Subsequently, we sample to generate scale-imbalanced data, with an imbalance ratio of up to 12:1. Our model is implemented in PyTorch and trained on a 40GB NVIDIA A100 GPU. We use the Adam optimizer (Kingma 2014) with a learning rate of 0.001. All experiments are repeated for 10 runs, each with 1000 epochs, and early stopping is applied. For evaluation, we adopt the widely used Macro-F1 and Micro-F1 metrics in graph classification tasks (Wang et al. 2022). The hyperparameter η is set to 0.001. The search space for the number of distilled tokens K_v in the node-level compact token bank and K_g in the graph-level distilled token bank is set to $[1, 128]$, and the Top-K selection threshold for graph-level compact tokens is searched in the range $[1, K_g]$. ***In our experiments, we use GIN as the backbone; other model also use GIN as encoder, whenever applicable, to ensure fair comparison.***

Experimental Result Analysis

Due to space limitations, we present the experimental results for six datasets. The remaining experimental results are provided in the **Supplement**.

► **Scale-imbalance Graph Classification.** Table 2 presents the overall performance, while Table 4 reports the fine-grained performance on large-sacle and small-sacle graphs. Specifically, GIN and most graph kernel methods perform the worst on scale-imbalanced datasets, exposing their limitations. Graph pooling methods (e.g., DGCNN and DiffPool) achieve better performance than GIN in classifying small-sacle graphs. GIN$^+$ incorporates techniques such as edge feature aggregation and residual connections, resulting in significantly improved accuracy. Graph Transformer (GTs)

Model	PROTEINS		NCI1		PTC-MR		REDDIT-B		COLLAB		IMDB-MULTI	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
GK	$55.41_{\pm 2.30}$	$60.77_{\pm 1.64}$	$63.19_{\pm 0.76}$	$63.33_{\pm 0.57}$	$36.84_{\pm 7.50}$	$51.56_{\pm 5.59}$	$68.48_{\pm 12.26}$	$71.74_{\pm 10.46}$	$28.56_{\pm 7.97}$	$32.24_{\pm 7.95}$	$27.09_{\pm 6.82}$	$33.68_{\pm 3.02}$
WL	$58.62_{\pm 2.24}$	$59.89_{\pm 2.25}$	$44.20_{\pm 10.87}$	$53.63_{\pm 4.07}$	$35.69_{\pm 8.32}$	$50.24_{\pm 9.27}$	$65.29_{\pm 4.37}$	$66.33_{\pm 5.61}$	$14.95_{\pm 6.20}$	$30.80_{\pm 15.87}$	$16.60_{\pm 0.20}$	$33.16_{\pm 0.53}$
GIN	$53.48_{\pm 2.03}$	$58.00_{\pm 4.19}$	$61.60_{\pm 2.20}$	$61.84_{\pm 2.29}$	$34.56_{\pm 6.32}$	$48.41_{\pm 7.07}$	$66.60_{\pm 2.27}$	$67.41_{\pm 2.23}$	$62.83_{\pm 1.41}$	$64.10_{\pm 1.53}$	$20.80_{\pm 4.91}$	$34.73_{\pm 2.16}$
CurGraph	$41.89_{\pm 6.97}$	$62.13_{\pm 1.25}$	$63.24_{\pm 2.28}$	$59.78_{\pm 1.37}$	$35.76_{\pm 8.13}$	$43.44_{\pm 3.46}$	$64.33_{\pm 2.35}$	$65.75_{\pm 4.10}$	$24.11_{\pm 0.85}$	$56.65_{\pm 0.11}$	$16.50_{\pm 0.31}$	$32.89_{\pm 1.25}$
Mixup	$63.36_{\pm 6.28}$	$65.57_{\pm 4.33}$	$49.49_{\pm 16.16}$	$57.81_{\pm 7.88}$	$32.55_{\pm 5.75}$	$45.17_{\pm 7.24}$	$63.64_{\pm 0.12}$	$67.99_{\pm 0.07}$	$21.58_{\pm 5.50}$	$49.66_{\pm 14.78}$	$16.67_{\pm 0.20}$	$33.17_{\pm 0.62}$
DGCNN	$66.56_{\pm 2.31}$	$71.43_{\pm 1.73}$	$60.39_{\pm 2.53}$	$61.74_{\pm 1.40}$	$41.30_{\pm 9.49}$	$46.84_{\pm 6.01}$	$58.47_{\pm 12.77}$	$63.57_{\pm 7.67}$	$34.98_{\pm 17.93}$	$47.29_{\pm 15.60}$	$32.21_{\pm 6.97}$	$37.62_{\pm 2.62}$
DiffPool	$68.38_{\pm 4.20}$	$72.05_{\pm 3.61}$	$66.97_{\pm 1.83}$	$67.39_{\pm 1.66}$	$49.02_{\pm 4.57}$	$50.56_{\pm 4.86}$	$63.47_{\pm 6.47}$	$63.75_{\pm 6.38}$	$49.85_{\pm 7.01}$	$50.96_{\pm 7.55}$	$28.55_{\pm 6.83}$	$36.36_{\pm 2.96}$
InfoGraph	$55.31_{\pm 3.30}$	$60.68_{\pm 3.52}$	$62.25_{\pm 1.53}$	$62.69_{\pm 2.01}$	$44.68_{\pm 8.14}$	$48.33_{\pm 4.91}$	$69.56_{\pm 3.46}$	$69.63_{\pm 4.28}$	$63.28_{\pm 1.68}$	$66.20_{\pm 1.33}$	$33.36_{\pm 1.60}$	$36.42_{\pm 1.10}$
GraphCL	$57.62_{\pm 5.50}$	$61.91_{\pm 3.84}$	$63.62_{\pm 5.43}$	$64.01_{\pm 4.75}$	$41.82_{\pm 6.67}$	$49.75_{\pm 5.25}$	$67.25_{\pm 7.92}$	$68.31_{\pm 6.36}$	$61.58_{\pm 4.16}$	$59.88_{\pm 4.01}$	$30.39_{\pm 2.53}$	$33.90_{\pm 1.57}$
GraphGPS	$65.54_{\pm 4.22}$	$69.26_{\pm 2.48}$	$62.96_{\pm 3.51}$	$64.02_{\pm 2.40}$	$44.14_{\pm 6.86}$	$49.97_{\pm 6.71}$	$66.16_{\pm 4.19}$	$68.42_{\pm 5.32}$	$24.11_{\pm 8.74}$	$56.65_{\pm 2.81}$	$16.87_{\pm 0.53}$	$33.49_{\pm 0.62}$
Exphormer	$64.01_{\pm 2.18}$	$67.33_{\pm 2.78}$	$65.16_{\pm 3.19}$	$65.23_{\pm 7.31}$	$46.45_{\pm 5.30}$	$50.66_{\pm 4.56}$	$66.48_{\pm 14.59}$	$67.81_{\pm 7.45}$	$21.03_{\pm 7.89}$	$36.71_{\pm 15.27}$	$25.52_{\pm 5.03}$	$33.60_{\pm 2.70}$
Graph-Mamba	$68.46_{\pm 3.91}$	$72.09_{\pm 3.16}$	$64.09_{\pm 2.82}$	$64.63_{\pm 2.14}$	$42.27_{\pm 5.86}$	$50.69_{\pm 6.99}$	$64.81_{\pm 12.47}$	$67.06_{\pm 8.99}$	$13.64_{\pm 5.97}$	$27.37_{\pm 14.82}$	$17.63_{\pm 3.07}$	$33.21_{\pm 0.52}$
GIN$^+$	$59.99_{\pm 3.50}$	$60.93_{\pm 3.65}$	$65.77_{\pm 2.80}$	$65.90_{\pm 2.76}$	$43.86_{\pm 7.23}$	$49.13_{\pm 5.15}$	$69.02_{\pm 1.05}$	$69.19_{\pm 1.10}$	$67.20_{\pm 1.99}$	$69.13_{\pm 2.13}$	$33.53_{\pm 0.45}$	$38.54_{\pm 0.61}$
TopoImb	$43.82_{\pm 11.74}$	$57.07_{\pm 13.48}$	$63.57_{\pm 1.69}$	$64.17_{\pm 1.33}$	$52.29_{\pm 2.09}$	$52.61_{\pm 2.18}$	$67.66_{\pm 3.93}$	$69.29_{\pm 3.47}$	$66.33_{\pm 1.38}$	$68.27_{\pm 1.69}$	$32.75_{\pm 5.19}$	$40.50_{\pm 2.97}$
ImbGNN	$66.87_{\pm 2.59}$	$67.07_{\pm 2.75}$	$61.74_{\pm 1.89}$	$61.92_{\pm 1.10}$	$43.28_{\pm 8.46}$	$52.94_{\pm 4.69}$	$64.55_{\pm 6.88}$	$66.49_{\pm 6.86}$	$49.23_{\pm 1.94}$	$50.22_{\pm 1.82}$	$22.21_{\pm 4.90}$	$33.18_{\pm 1.02}$
SOLT-GNN	$69.34_{\pm 3.05}$	$71.38_{\pm 3.34}$	$60.14_{\pm 2.38}$	$60.58_{\pm 2.04}$	$40.70_{\pm 3.27}$	$51.74_{\pm 5.25}$	$56.69_{\pm 8.70}$	$58.62_{\pm 9.02}$	$62.46_{\pm 4.80}$	$65.05_{\pm 5.77}$	$34.25_{\pm 2.71}$	$39.58_{\pm 2.00}$
Ours	$\mathbf{83.14}_{\pm \mathbf{2.08}}$	$\mathbf{83.42}_{\pm \mathbf{2.15}}$	$\mathbf{80.23}_{\pm \mathbf{3.77}}$	$\mathbf{80.36}_{\pm \mathbf{2.70}}$	$\mathbf{54.47}_{\pm \mathbf{0.45}}$	$\mathbf{56.08}_{\pm \mathbf{1.91}}$	$\mathbf{72.62}_{\pm \mathbf{4.03}}$	$\mathbf{73.94}_{\pm \mathbf{3.09}}$	$\mathbf{75.57}_{\pm \mathbf{2.76}}$	$\mathbf{80.45}_{\pm \mathbf{2.22}}$	$\mathbf{41.85}_{\pm \mathbf{2.28}}$	$\mathbf{45.05}_{\pm \mathbf{1.49}}$

Table 2: Performance comparison on **scale-imbalance** datasets. The best results are **highlighted** in bold, while the second-best results are underlined. We report the average and standard deviation over 10 runs.

Model	PROTEINS		NCI1		PTC-MR		REDDIT-B		COLLAB		IMDB-MULTI	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
GK	$56.70_{\pm 6.67}$	$59.68_{\pm 2.27}$	$59.37_{\pm 6.99}$	$62.66_{\pm 3.74}$	$43.91_{\pm 3.41}$	$46.88_{\pm 5.01}$	$66.60_{\pm 5.53}$	$67.94_{\pm 4.67}$	$59.20_{\pm 0.83}$	$61.07_{\pm 1.10}$	$23.20_{\pm 4.21}$	$34.86_{\pm 1.47}$
WL	$60.43_{\pm 1.40}$	$64.42_{\pm 0.79}$	$54.15_{\pm 7.71}$	$57.01_{\pm 3.36}$	$46.60_{\pm 4.22}$	$47.78_{\pm 3.99}$	$64.12_{\pm 3.21}$	$65.29_{\pm 4.17}$	$14.76_{\pm 7.31}$	$31.10_{\pm 18.90}$	$16.70_{\pm 0.24}$	$33.42_{\pm 0.65}$
GIN	$59.67_{\pm 1.87}$	$63.58_{\pm 1.76}$	$61.89_{\pm 5.08}$	$63.30_{\pm 3.19}$	$48.26_{\pm 3.11}$	$52.17_{\pm 4.36}$	$66.14_{\pm 1.90}$	$67.41_{\pm 2.32}$	$58.45_{\pm 1.80}$	$64.34_{\pm 3.55}$	$35.91_{\pm 5.03}$	$44.92_{\pm 3.66}$
CurGraph	$42.25_{\pm 6.55}$	$60.22_{\pm 6.54}$	$52.32_{\pm 1.43}$	$52.88_{\pm 1.62}$	$49.70_{\pm 1.43}$	$51.04_{\pm 1.65}$	$62.73_{\pm 3.58}$	$64.59_{\pm 4.18}$	$26.78_{\pm 2.12}$	$58.65_{\pm 3.31}$	$16.55_{\pm 0.15}$	$33.03_{\pm 0.39}$
Mixup	$56.60_{\pm 0.60}$	$62.23_{\pm 0.38}$	$42.42_{\pm 13.88}$	$54.09_{\pm 6.28}$	$44.43_{\pm 2.88}$	$46.25_{\pm 4.26}$	$65.40_{\pm 10.14}$	$67.01_{\pm 11.69}$	$24.48_{\pm 6.47}$	$56.65_{\pm 17.25}$	$16.64_{\pm 0.23}$	$33.29_{\pm 0.60}$
DGCNN	$58.59_{\pm 13.15}$	$62.27_{\pm 9.25}$	$46.21_{\pm 12.22}$	$53.77_{\pm 6.01}$	$45.15_{\pm 5.81}$	$49.62_{\pm 5.19}$	$46.08_{\pm 8.71}$	$46.90_{\pm 8.29}$	$43.74_{\pm 12.72}$	$50.27_{\pm 7.37}$	$22.12_{\pm 4.88}$	$33.52_{\pm 1.27}$
DiffPool	$55.65_{\pm 3.38}$	$56.90_{\pm 2.79}$	$65.01_{\pm 1.97}$	$77.71_{\pm 1.36}$	$46.94_{\pm 4.03}$	$47.92_{\pm 3.32}$	$56.48_{\pm 13.94}$	$61.79_{\pm 9.55}$	$28.32_{\pm 6.52}$	$28.50_{\pm 6.07}$	$29.29_{\pm 3.29}$	$35.41_{\pm 2.89}$
InfoGraph	$61.18_{\pm 2.22}$	$64.27_{\pm 2.58}$	$63.44_{\pm 4.42}$	$64.69_{\pm 4.10}$	$47.19_{\pm 3.73}$	$51.07_{\pm 2.84}$	$71.31_{\pm 5.94}$	$72.22_{\pm 6.73}$	$60.03_{\pm 4.26}$	$63.75_{\pm 4.82}$	$37.22_{\pm 6.35}$	$41.83_{\pm 7.01}$
GraphCL	$63.72_{\pm 3.03}$	$65.08_{\pm 4.29}$	$65.08_{\pm 3.27}$	$66.82_{\pm 3.19}$	$48.57_{\pm 4.23}$	$52.90_{\pm 4.65}$	$70.48_{\pm 4.28}$	$71.75_{\pm 5.35}$	$58.47_{\pm 7.88}$	$62.82_{\pm 6.38}$	$36.98_{\pm 5.16}$	$40.81_{\pm 6.48}$
GraphGPS	$65.95_{\pm 2.46}$	$66.96_{\pm 3.23}$	$63.20_{\pm 0.94}$	$63.46_{\pm 0.81}$	$49.81_{\pm 2.95}$	$50.07_{\pm 2.88}$	$66.42_{\pm 3.21}$	$67.95_{\pm 4.17}$	$13.18_{\pm 8.47}$	$28.13_{\pm 22.09}$	$16.90_{\pm 0.20}$	$33.95_{\pm 0.53}$
Exphormer	$51.77_{\pm 6.39}$	$54.73_{\pm 5.82}$	$49.17_{\pm 8.89}$	$54.77_{\pm 3.48}$	$44.99_{\pm 3.29}$	$45.97_{\pm 2.60}$	$56.98_{\pm 6.23}$	$57.25_{\pm 6.24}$	$15.06_{\pm 4.85}$	$20.75_{\pm 8.44}$	$21.11_{\pm 4.48}$	$33.03_{\pm 1.93}$
Graph-Mamba	$60.16_{\pm 2.54}$	$60.88_{\pm 2.20}$	$64.49_{\pm 1.92}$	$64.70_{\pm 1.70}$	$47.85_{\pm 3.22}$	$48.19_{\pm 2.91}$	$67.45_{\pm 10.88}$	$68.18_{\pm 5.78}$	$28.65_{\pm 17.81}$	$31.96_{\pm 16.16}$	$16.61_{\pm 0.20}$	$33.16_{\pm 0.53}$
GIN$^+$	$55.23_{\pm 4.02}$	$56.32_{\pm 4.96}$	$66.27_{\pm 1.20}$	$66.33_{\pm 1.14}$	$50.20_{\pm 3.44}$	$52.12_{\pm 4.27}$	$65.03_{\pm 3.64}$	$66.08_{\pm 3.80}$	$58.38_{\pm 1.37}$	$59.75_{\pm 1.44}$	$40.94_{\pm 2.74}$	$45.93_{\pm 2.86}$
TopoImb	$59.94_{\pm 4.32}$	$61.19_{\pm 4.61}$	$62.41_{\pm 0.66}$	$62.46_{\pm 0.66}$	$48.83_{\pm 2.90}$	$49.71_{\pm 1.98}$	$69.15_{\pm 3.83}$	$69.47_{\pm 3.70}$	$60.91_{\pm 2.15}$	$62.71_{\pm 2.42}$	$39.09_{\pm 2.72}$	$45.25_{\pm 1.64}$
ImbGNN	$62.41_{\pm 11.37}$	$64.33_{\pm 4.65}$	$62.01_{\pm 0.94}$	$62.09_{\pm 0.99}$	$48.72_{\pm 4.22}$	$50.33_{\pm 3.26}$	$54.47_{\pm 9.62}$	$58.47_{\pm 8.74}$	$53.42_{\pm 0.74}$	$55.23_{\pm 1.01}$	$25.97_{\pm 5.55}$	$37.24_{\pm 2.85}$
SOLT-GNN	$64.60_{\pm 4.66}$	$65.56_{\pm 4.83}$	$58.75_{\pm 0.73}$	$59.59_{\pm 0.77}$	$43.63_{\pm 1.61}$	$47.54_{\pm 4.33}$	$49.17_{\pm 6.76}$	$49.57_{\pm 6.78}$	$59.76_{\pm 1.81}$	$61.79_{\pm 2.09}$	$37.59_{\pm 3.24}$	$40.43_{\pm 3.27}$
Ours	$\mathbf{73.36}_{\pm \mathbf{2.71}}$	$\mathbf{75.92}_{\pm \mathbf{1.30}}$	$\mathbf{72.78}_{\pm \mathbf{1.50}}$	$\mathbf{73.45}_{\pm \mathbf{2.12}}$	$\mathbf{52.28}_{\pm \mathbf{1.86}}$	$\mathbf{53.30}_{\pm \mathbf{1.56}}$	$\mathbf{75.63}_{\pm \mathbf{2.02}}$	$\mathbf{75.69}_{\pm \mathbf{2.56}}$	$\mathbf{63.54}_{\pm \mathbf{2.81}}$	$\mathbf{78.03}_{\pm \mathbf{1.26}}$	$\mathbf{46.46}_{\pm \mathbf{2.08}}$	$\mathbf{48.04}_{\pm \mathbf{1.58}}$

Table 3: Performance comparison on **scale-balance** datasets. The best results are **highlighted** in bold, while the second-best results are underlined. We report the average and standard deviation over 10 runs.

models (such as GraphGPS, Exphormer, and Graph-Mamba) further enhance GIN with global attention mechanisms, thus alleviating the small-sacle graph issue. Counterintuitively, topology-specific models do not achieve dominant performance. For instance, TopoImb and ImbGNN improve graph representations through topology-aware modeling and enhanced G^2G, respectively. Furthermore, we report the performance on large-scale and small-scale graphs in Table 4. It can be observed that these topology-imbalance methods fail to achieve the expected results on small-scale graphs, and their performance on large-scale graphs is generally inferior to that on small-scale graphs. In contrast, U2B achieves dominating performance by balancing learning at both the node and graph levels and integrating consistency loss, with performance improvements of up to **22.19%**.

► **Scale-balance Graph Classification.** As shown in Table 3, in scale-balanced scenarios, models designed for topology-imbalanced learning perform suboptimally compared with conventional models such as GIN, indicating that their strategies may be ineffective in this setting. Graph contrastive learning methods, such as InfoGraph and GraphCL, improve performance through mutual information consistency and graph augmentation. Although CurGraph leverages curriculum learning, it exhibits instability on certain datasets. GIN$^+$ integrates techniques like residual connections and positional encodings, outperforming most baselines. Graph Transformer models excel on binary classification tasks, but their performance is less impressive on balanced multi-class datasets. In this scenario, U2B still achieves competitive accuracy, as it effectively extracts base features that aid the model in classification. Compared with the strongest baseline, our method achieves performance improvements of up to **21.28%**.

Scalability Analysis of U2B

To demonstrate the scalability of U2B, we further evaluated it in scale-imbalanced scenarios by employing different GNN variants and Graph Transformers as backbone. The results (as shown in Table 5) indicate that all backbone variants achieved significant performance improvements when integrated with U2B. These findings validate the strong scalability of our approach, which can substantially enhance the performance of conventional graph classification models.

Model	PROTEINS		NCI1		IMDB-MULTI	
	Small-scale	Large-scale	Small-scale	Large-scale	Small-scale	Large-scale
GIN	$56.12_{\pm 3.18}$	$64.28_{\pm 3.52}$	$58.39_{\pm 3.12}$	$70.88_{\pm 6.01}$	$32.19_{\pm 4.41}$	$33.98_{\pm 5.06}$
CurGraph	$54.70_{\pm 2.22}$	$78.62_{\pm 2.25}$	$51.20_{\pm 4.34}$	$68.07_{\pm 6.25}$	$31.28_{\pm 0.97}$	$38.06_{\pm 3.11}$
DGCNN	$66.28_{\pm 2.25}$	$81.96_{\pm 0.63}$	$56.85_{\pm 1.78}$	$78.75_{\pm 0.12}$	$37.48_{\pm 3.19}$	$38.10_{\pm 4.84}$
DiffPool	$68.67_{\pm 3.40}$	$83.65_{\pm 6.20}$	$65.93_{\pm 2.22}$	$72.46_{\pm 6.42}$	$32.35_{\pm 6.97}$	$37.61_{\pm 3.52}$
InfoGraph	$58.42_{\pm 2.67}$	$65.10_{\pm 3.19}$	$60.76_{\pm 3.29}$	$70.64_{\pm 5.17}$	$34.22_{\pm 4.61}$	$37.29_{\pm 2.13}$
GraphCL	$60.00_{\pm 2.28}$	$65.88_{\pm 3.41}$	$61.33_{\pm 4.67}$	$74.75_{\pm 5.02}$	$31.24_{\pm 5.94}$	$35.98_{\pm 5.19}$
GraphGPS	$66.02_{\pm 2.85}$	$80.40_{\pm 6.46}$	$62.81_{\pm 4.77}$	$68.18_{\pm 7.61}$	$34.14_{\pm 1.10}$	$31.38_{\pm 3.01}$
Exphormer	$65.42_{\pm 2.84}$	$73.91_{\pm 9.07}$	$63.96_{\pm 12.87}$	$68.83_{\pm 2.48}$	$33.59_{\pm 3.45}$	$33.63_{\pm 3.79}$
Graph-Mamba	$68.26_{\pm 3.61}$	$85.25_{\pm 3.52}$	$63.49_{\pm 3.60}$	$68.60_{\pm 6.11}$	$30.45_{\pm 3.19}$	$34.07_{\pm 1.08}$
GIN+	$60.32_{\pm 3.19}$	$63.02_{\pm 7.43}$	$65.92_{\pm 2.34}$	$65.78_{\pm 8.22}$	$35.99_{\pm 0.65}$	$45.03_{\pm 1.97}$
TopoImb	$55.36_{\pm 6.68}$	$62.87_{\pm 36.75}$	$61.72_{\pm 1.24}$	$72.92_{\pm 1.57}$	$38.40_{\pm 2.56}$	$\mathbf{47.62}_{\pm \mathbf{2.74}}$
ImbGNN	$68.30_{\pm 0.88}$	$63.23_{\pm 3.07}$	$58.68_{\pm 1.86}$	$70.89_{\pm 1.56}$	$32.76_{\pm 6.13}$	$34.07_{\pm 7.33}$
SOLT-GNN	$67.34_{\pm 3.57}$	$83.88_{\pm 3.64}$	$58.52_{\pm 1.91}$	$68.77_{\pm 7.96}$	$36.64_{\pm 1.39}$	$42.24_{\pm 2.76}$
Ours	$\mathbf{82.34}_{\pm \mathbf{1.92}}$	$\mathbf{86.31}_{\pm \mathbf{2.68}}$	$\mathbf{79.68}_{\pm \mathbf{2.30}}$	$\mathbf{81.42}_{\pm \mathbf{2.05}}$	$\mathbf{43.48}_{\pm \mathbf{1.81}}$	$45.54_{\pm 2.15}$

Table 4: Micro-F1 Performance comparison on **Scale-imbalance** datasets.

Method	PROTEINS		NCI1		IMDB-MULTI	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
GIN	$53.48_{\pm 2.03}$	$58.00_{\pm 4.19}$	$61.60_{\pm 2.20}$	$61.84_{\pm 2.29}$	$20.80_{\pm 4.91}$	$34.73_{\pm 2.16}$
GIN + U2B	$\mathbf{83.14}_{\pm \mathbf{2.08}}$	$\mathbf{83.42}_{\pm \mathbf{2.15}}$	$\mathbf{80.23}_{\pm \mathbf{3.77}}$	$\mathbf{80.36}_{\pm \mathbf{2.70}}$	$\mathbf{41.85}_{\pm \mathbf{2.28}}$	$\mathbf{45.05}_{\pm \mathbf{1.49}}$
GCN	$51.29_{\pm 1.10}$	$57.02_{\pm 5.02}$	$60.39_{\pm 1.21}$	$60.48_{\pm 1.17}$	$20.47_{\pm 4.13}$	$32.93_{\pm 1.35}$
GCN + U2B	$\mathbf{74.65}_{\pm \mathbf{2.49}}$	$\mathbf{76.44}_{\pm \mathbf{2.61}}$	$\mathbf{74.72}_{\pm \mathbf{6.89}}$	$\mathbf{74.73}_{\pm \mathbf{1.87}}$	$\mathbf{35.36}_{\pm \mathbf{2.73}}$	$\mathbf{40.25}_{\pm \mathbf{1.03}}$
GraphSAGE	$52.95_{\pm 1.10}$	$59.18_{\pm 1.17}$	$61.24_{\pm 0.66}$	$61.63_{\pm 0.74}$	$19.03_{\pm 2.10}$	$31.42_{\pm 1.12}$
GraphSAGE + U2B	$\mathbf{78.43}_{\pm \mathbf{1.67}}$	$\mathbf{78.46}_{\pm \mathbf{1.33}}$	$\mathbf{75.19}_{\pm \mathbf{4.43}}$	$\mathbf{75.82}_{\pm \mathbf{4.38}}$	$\mathbf{34.29}_{\pm \mathbf{3.10}}$	$\mathbf{39.46}_{\pm \mathbf{2.38}}$
Exphormer	$64.01_{\pm 2.18}$	$67.33_{\pm 2.78}$	$65.16_{\pm 3.19}$	$65.23_{\pm 7.31}$	$25.52_{\pm 5.03}$	$33.60_{\pm 2.70}$
Exphormer + U2B	$\mathbf{80.21}_{\pm \mathbf{3.42}}$	$\mathbf{80.28}_{\pm \mathbf{2.02}}$	$\mathbf{78.48}_{\pm \mathbf{2.19}}$	$\mathbf{78.61}_{\pm \mathbf{3.20}}$	$\mathbf{39.61}_{\pm \mathbf{2.06}}$	$\mathbf{43.25}_{\pm \mathbf{2.81}}$
GraphMamba	$68.46_{\pm 3.91}$	$72.09_{\pm 3.16}$	$64.09_{\pm 2.82}$	$64.63_{\pm 2.14}$	$17.63_{\pm 3.07}$	$33.21_{\pm 0.52}$
GraphMamba + U2B	$\mathbf{86.71}_{\pm \mathbf{3.04}}$	$\mathbf{86.93}_{\pm \mathbf{4.15}}$	$\mathbf{82.29}_{\pm \mathbf{2.47}}$	$\mathbf{82.45}_{\pm \mathbf{2.54}}$	$\mathbf{32.07}_{\pm \mathbf{5.62}}$	$\mathbf{38.82}_{\pm \mathbf{4.97}}$

Table 5: **Scale-imbalance** graph classification performance with other models as backbone.

Ablation Study

To evaluate the contribution of each component in U2B, we create six variants: ❶ *w/o NC* means that the Node-level Converter is removed; ❷ *w/o GC* means that the Graph-level Converter is removed; ❸ *w/o Sink* means that the a Sinkhorn Knopp-based static Constraint in the Node-level Converter is removed; ❹ *w/o Dyna* means that the Dyanmic Constraint in the Graph-level Converter is removed; ❺ *w/o ConA* means that the Consistency Alignment Loss between graph representation and node representations is removed.

From Figure 4, all variants perform worse than U2B. The w/o NC variant has the lowest accuracy on PROTEINS datatset, showing that the node-level Converter improves small-scale graph expressiveness by capturing key node patterns. The w/o GC variant also underperforms, highlighting the effectiveness of graph-level Converter in enriching graph representations. Additionally, the constraint regularization terms (w/o Sink and w/o ConA) are crucial for balancing graphs of different scales.

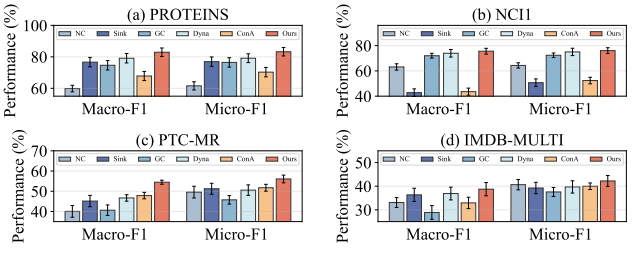


Figure 4: Ablation Experiments on Scale-imbalance datasets.

Parameters Sensitivity Analysis

We evaluate the sensitivity of two key hyperparameters in the U2B framework : ❶ the number of node-level distilled tokens K_v and ❷ the number of graph-level distilled tokens K_g. Specifically, using four datasets under scale-imbalance scenario, Figure 5 illustrate the performance impact of The optimal range for K_g and K_v is between [16, 64]. When the number of tokens is too large, the model may struggle to focus on extracting representative base features.

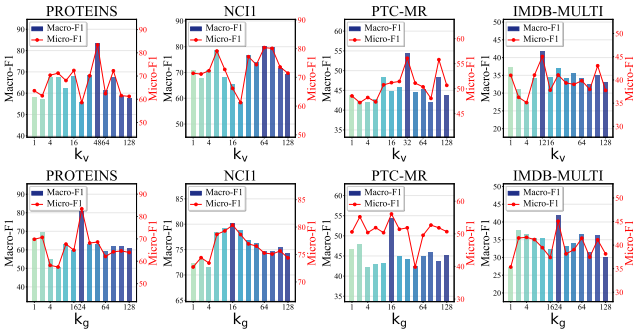
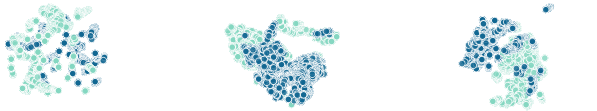


Figure 5: Impact of K_v and K_g on Scale-imbalance datasets.

Visualization of graph representations

As shown in Figure 6, we further visualized the final graph representations used for classification, using the scale-imbalanced NCI1 dataset as an example to analyze the quality of the representations. The initial GIN representations were disorganized. As we gradually incorporated node-level and graph-level Converters into GIN, the graph representations become highly discriminative. This demonstrates that U2B can effectively balance graph representations of different scales and enhance performance.



(a) GIN (b) + *Node-level Converter* (c) + *Graph-level Converter*

Figure 6: Representation visualization on NCI1 dataset.

6 Conclusion

In this paper, we develop U2B to tackle the challenges posed by comprehensive scale distribution scenarios in graph classification. U2B employs a novel distillation-refinement framework to extract fundamental features at both the node and graph levels. By introducing static constraints at the node-level converter and dynamic constraints at the graph-level converter, U2B reduces bias toward large-scale graphs while maintaining performance on samll-scale graphs. Finally, we introduce a consistency alignment loss to further enhance the coherence between node-level and graph-level representations. Extensive experiments conducted on 8 benchmark datasets and 16 baselines demonstrate that U2B achieves optimal accuracy, with improvements of up to 22.19%.

References

Asano, Y. M.; Rupprecht, C.; and Vedaldi, A. 2019. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*.

Cai, D.; and Lam, W. 2020. Graph transformer for graph-to-sequence learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 7464–7471.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, 132–149.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.

Hamilton, W. L. 2020. *Graph representation learning*. Morgan & Claypool Publishers.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Liu, Z.; Mao, Q.; Liu, C.; Fang, Y.; and Sun, J. 2022. On size-oriented long-tailed graph classification of graph neural networks. In *Proceedings of the ACM web conference 2022*, 1506–1516.

Luo, Y.; Shi, L.; and Wu, X.-M. 2025. Can Classic GNNs Be Strong Baselines for Graph-level Tasks? Simple Architectures Meet Excellence. *arXiv preprint arXiv:2502.09263*.

Noutahi, E.; Beaini, D.; Horwood, J.; Giguère, S.; and Tossou, P. 2019. Towards interpretable sparse graph representation learning with laplacian pooling. *arXiv preprint arXiv:1905.11577*.

Nowozin, S.; Cseke, B.; and Tomioka, R. 2016. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29.

Qin, J.; Yuan, H.; Sun, Q.; Xu, L.; Yuan, J.; Huang, P.; Wang, Z.; Fu, X.; Peng, H.; Li, J.; et al. 2024. Igl-bench: Establishing the comprehensive benchmark for imbalanced graph learning. *arXiv preprint arXiv:2406.09870*.

Qu, L.; Zhu, H.; Zheng, R.; Shi, Y.; and Yin, H. 2021. Imgagn: Imbalanced network embedding via generative adversarial graph networks. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 1390–1398.

Rampášek, L.; Galkin, M.; Dwivedi, V. P.; Luu, A. T.; Wolf, G.; and Beaini, D. 2022. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35: 14501–14515.

Shervashidze, N.; Schweitzer, P.; Van Leeuwen, E. J.; Mehlhorn, K.; and Borgwardt, K. M. 2011. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9).

Shervashidze, N.; Vishwanathan, S.; Petri, T.; Mehlhorn, K.; and Borgwardt, K. 2009. Efficient graphlet kernels for large graph comparison. In *Artificial intelligence and statistics*, 488–495. PMLR.

Shi, M.; Tang, Y.; Zhu, X.; Wilson, D.; and Liu, J. 2020. Multi-class imbalanced graph convolutional network learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*.

Shirzad, H.; Velingker, A.; Venkatachalam, B.; Sutherland, D. J.; and Sinop, A. K. 2023. Exphormer: Sparse transformers for graphs. In *International Conference on Machine Learning*, 31613–31632. PMLR.

Sun, F.-Y.; Hoffmann, J.; Verma, V.; and Tang, J. 2019. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000*.

Verdier, H.; Duval, M.; Laurent, F.; Cassé, A.; Vestergaard, C. L.; and Masson, J.-B. 2021. Learning physical properties of anomalous random walks using graph neural networks. *Journal of Physics A: Mathematical and Theoretical*, 54(23): 234001.

Wang, Y.; Wang, W.; Liang, Y.; Cai, Y.; and Hooi, B. 2021a. Curgraph: Curriculum learning for graph classification. In *Proceedings of the web conference 2021*, 1238–1248.

Wang, Y.; Wang, W.; Liang, Y.; Cai, Y.; and Hooi, B. 2021b. Mixup for node and graph classification. In *Proceedings of the Web Conference 2021*, 3663–3674.

Wang, Y.; Xu, Y.; Yang, J.; Wu, M.; Li, X.; Xie, L.; and Chen, Z. 2024. Graph-aware contrasting for multivariate time-series classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 15725–15734.

Wang, Y.; Zhao, Y.; Shah, N.; and Derr, T. 2022. Imbalanced graph classification via graph-of-graph neural networks. In *Proceedings of the 31st ACM international conference on information & knowledge management*, 2067–2076.

Wei, L.; Zhao, H.; He, Z.; and Yao, Q. 2023. Neural architecture search for GNN-based graph classification. *ACM Transactions on Information Systems*, 42(1): 1–29.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Yu, P. S. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1): 4–24.

Xie, Y.; Liang, Y.; Gong, M.; Qin, A. K.; Ong, Y.-S.; and He, T. 2022. Semisupervised graph neural networks for graph classification. *IEEE Transactions on Cybernetics*, 53(10): 6222–6235.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Xu, W.; Wang, P.; Zhao, Z.; Wang, B.; Wang, X.; and Wang, Y. 2024. When imbalance meets imbalance: Structure-driven learning for imbalanced graph classification. In *Proceedings of the ACM Web Conference 2024*, 905–913.

Ying, Z.; You, J.; Morris, C.; Ren, X.; Hamilton, W.; and Leskovec, J. 2018. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 31.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823.

Yun, S.; Jeong, M.; Kim, R.; Kang, J.; and Kim, H. J. 2019. Graph transformer networks. *Advances in neural information processing systems*, 32.

Zeng, L.; Li, L.; Gao, Z.; Zhao, P.; and Li, J. 2023. Imgcl: Revisiting graph contrastive learning on imbalanced node classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11138–11146.

Zhang, M.; Cui, Z.; Neumann, M.; and Chen, Y. 2018. An end-to-end deep learning architecture for graph classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhao, T.; Luo, D.; Zhang, X.; and Wang, S. 2022. Topoimb: Toward topology-level imbalance in learning from graphs. In *Learning on Graphs Conference*, 37–1. PMLR.

Zhao, T.; Zhang, X.; and Wang, S. 2021. Graphsmote: Imbalanced node classification on graphs with graph neural networks. In *Proceedings of the 14th ACM international conference on web search and data mining*, 833–841.

Reproducibility Checklist

Instructions for Authors:

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "Type your response here" text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "Type your response here" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

---

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) Type your response here

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) Type your response here

2.4. Proofs of all novel claims are included (yes/partial/no) Type your response here

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) Type your response here

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) Type your response here

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) Type your response here

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) Type your response here

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) yes

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) NA

3.5. All datasets drawn from the existing literature (po-

tentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) yes

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) yes

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) yes

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes