RARE: Robust Masked Graph Autoencoder

Wenxuan Tu, Qing Liao, Sihang Zhou, Xin Peng, Chuan Ma, *Member, IEEE*, Zhe Liu, *Senior Member, IEEE*, Xinwang Liu, *Senior Member, IEEE*, Zhiping Cai, and Kunlun He

***Abstract*—Masked graph autoencoder (MGAE) has emerged as a promising self-supervised graph pre-training (SGP) paradigm due to its simplicity and effectiveness. However, existing efforts perform the mask-then-reconstruct operation in the raw data space as is done in computer vision (CV) and natural language processing (NLP) areas, while neglecting the important non-Euclidean property of graph data. As a result, the highly unstable local structures largely increase the uncertainty in inferring masked data and decrease the reliability of the exploited self-supervision signals, leading to inferior representations for downstream evaluations. To address this issue, we propose a novel SGP method termed Robust mAsked gRaph autoEncoder (RARE) to improve the certainty in inferring masked data and the reliability of the self-supervision mechanism by further masking and reconstructing node samples in the high-order latent feature space. Through both theoretical and empirical analyses, we have discovered that performing a joint mask-then-reconstruct strategy in both latent feature and raw data spaces could yield improved stability and performance. To this end, we elaborately design a masked latent feature completion scheme, which predicts latent features of masked nodes under the guidance of high-order sample correlations that are hard to be observed from the raw data perspective. Specifically, we first adopt a latent feature predictor to predict the masked latent features from the visible ones. Next, we encode the raw data of masked samples with a momentum graph encoder and subsequently employ the resulting representations to improve the predicted results through latent feature matching. Extensive experiments on seventeen datasets have demonstrated the effectiveness and robustness of RARE against state-of-the-art (SOTA) competitors across three downstream tasks.**

***Index Terms*—Incomplete multi-view learning, classification, masked graph autoencoder, robustness.**

I. Introduction

MASKED autoencoders (MAEs) have emerged as the dominant technique for self-supervised vision and language pre-training. The objective of MAEs is to learn generalized sample representations from massive unlabeled data by recovering partially masked content (e.g., image patches or word embeddings) from observations. Due to their simplicity and powerful local structure modeling capabilities, advanced efforts in this field [1], [2], [3] have garnered significant interest among researchers. These methods have demonstrated impressive performance across a wide range of real-world applications, including medical image analysis [4], natural language understanding [5], and 3D object detection [6].

Recent studies have shown that applying MAEs to facilitate graph machine learning has become a topic of increasing interest. The success of masked graph autoencoders (MGAEs) lies in the mask-then-reconstruct operation. In this setup, a portion of visible nodes or edges are randomly masked and then adopted as self-supervision signals to guide the model learning, so as to allow the network to explore the underlying structural information for downstream evaluations [7], [8], [9], [10], [11], [12], [13]. Despite their promising performance on various graph-oriented tasks [14], [15], [16], existing MGAEs overlook the inherent distinction between images (or texts) and graphs, i.e., images (or texts) are Euclidean while graphs are non-Euclidean. In other words, the nearby structure of an image patch or a sub-sentence has higher semantic certainty and is more stable than that of a sub-graph. Therefore, the quality of the self-supervision guidance provided by a masked image patch or a masked word embedding would be much higher than that provided by a masked node (or edge). Specifically, as shown in Fig. 1(a), since the relative spatial distribution of organs on a dog is quite certain, people can easily imagine the masked image patches based on the observed content within an incomplete dog photo. Similarly, in Fig. 1(b), the strong context correlation among words helps us involuntarily fill an incomplete sentence with meaningful words. Comparatively, in a social network where nodes are entities and edges are interactions, the neighborhood structure of an entity within a graph varies a lot, as shown in Fig. 1(c). When a node or an edge

Manuscript received 2 April 2023; revised 30 August 2023; accepted 11 November 2023. Date of publication 23 November 2023; date of current version 4 October 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 62325604, 62276271, 62006237, and 62002170, in part by the Research Initiation Project of Zhejiang Lab under Grant 2022PD0AC02, in part by the Youth Foundation Project of Zhejiang Lab under Grant K2023PD0AA01, and in part by the Science and Technology Innovation Program of Hunan Province under Grant 2022RC3061. Recommended for acceptance by Z. Guan. *(Corresponding authors: Kunlun He; Qing Liao; Xinwang Liu.)*

Wenxuan Tu, Xin Peng, Xinwang Liu, and Zhiping Cai are with the College of Computer, National University of Defense Technology, Changsha 410073, China (e-mail: wenxuantu@163.com; pengxin3252@163.com; xinwangliu@nudt.edu.cn; zpcai@nudt.edu.cn).

Qing Liao is with the College of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: liaoqing@hit.edu.cn).

Sihang Zhou is with the College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China (e-mail: sihangjoe@gmail.com).

Chuan Ma is with Zhejiang Lab, Hangzhou 311121, China, and also with the Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, Nanjing 211189, China (e-mail: chuan.ma@zhejianglab.edu.cn).

Zhe Liu is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China, and also with Zhejiang Lab, Hangzhou 311121, China (e-mail: zhe.liu@nuaa.edu.cn).

Kunlun He is with Medical Big Data Research Center, Chinese PLA General Hospital, Beijing 100853, China (e-mail: kunlunhe@plagh.org).

This article has supplementary downloadable material available at https://doi.org/10.1109/TKDE.2023.3335222, provided by the authors.

Digital Object Identifier 10.1109/TKDE.2023.3335222

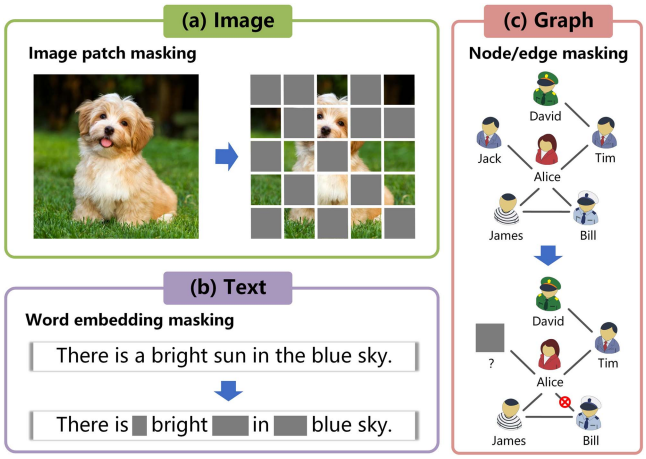

Fig. 1. Information masking on different types of data. For instance, masking image patches on images (a) or word embeddings on texts (b) would not alter the underlying semantics of the original data to some extent. In other words, even after a portion of image patches or word embeddings are masked, people can still recognize the object visually or understand the language content by inferring the invisible content based on observed contexts. However, masking nodes or edges on non-Euclidean graphs (c) may unexpectedly increase the uncertainty when inferring masked data, since the nearby structure of a node or an edge is less stable and has lower certainty than that of an image or a text.

is entirely masked, it is hard to identify the removed entity or ascertain whether two entities keep in contact directly, since it is common for two unconnected entities to have conjoint neighbors or valuable higher-order relationships that are hard to be observed from the raw data perspective in a real-world social graph. Consequently, although in most cases, the mask-then-reconstruct principle is effective for learning valuable node representations, directly recovering the masked nodes and edges driven by the low-level raw data would put the corresponding model at risk of being confused by the local structural ambiguity. Based on these observations, we argue that the non-Euclidean property of graph data could to some extent trigger uncertainty in inferring masked data and may negatively affect the reliability of the self-supervision mechanism. In this circumstance, the robustness of the model may be compromised when applying a masked autoencoder to process graphs straightforwardly.

To address the above issue, we propose a novel method termed **R**obust m**A**sked g**R**aph auto**E**ncoder (RARE) for self-supervised graph pre-training. The main idea of RARE is to integrate implicit and explicit self-supervision mechanisms for masked content recovery by performing a joint mask-then-reconstruct strategy in both latent feature and raw data spaces. The effectiveness of our method lies in the fact that unlike the model optimization driven by the low-level raw data only [8], [9], the self-supervision mechanism of RARE could be further enhanced by incorporating more informative high-order sample correlations that are hard to be observed from the raw data perspective [17], [18], [19]. To this end, we design a masked latent feature completion scheme that includes two steps. Specifically, we first adopt a latent feature predictor to assist the graph encoder in extracting more compressed features by predicting the latent features of masked samples based on observations. To further enhance the integrity and accuracy of predicted representations, we encode the raw data of masked samples using a momentum graph encoder and leverage the resultant representations to guide the latent feature prediction through information matching. With such persistent and informative signals as self-supervision guidance, each masked sample in the latent space is encouraged to explore reliable information from available features. As the implicit self-supervision signals become more reliable and the reconstructed content becomes more accurate, the model is enforced to promote greater information encoding capability, thereby generating more robust and generalized node representations for downstream evaluations. The main contributions of this work are summarized below:

- We propose a novel SGP framework termed RARE to enhance the robustness of masked graph autoencoders. It not only eases the instability of the self-supervision mechanism driven by the non-Euclidean raw graph data, but also achieves a good accuracy-efficiency trade-off.
- We incorporate a simple but effective masked latent feature completion scheme into a masked graph autoencoder. This design can enhance the certainty in inferring masked data and the reliability of the self-supervision mechanism by exploiting more informative high-order sample correlations to drive the model learning.
- Extensive experiments on seventeen datasets across three downstream tasks demonstrate the effectiveness and robustness of RARE against competitors. Moreover, a series of elaborate ablation studies also verify that RARE can indeed unleash the full potential of MAEs to provide a comprehensive understanding of graphs.

The remainder of this paper is organized as follows. In Section II, we review related work in areas of self-supervised graph pre-training, masked graph autoencoders, and self-distillation on graphs. Section III presents defined notations, the network designs, loss functions, and theoretical discussions. In Section IV, we conduct experiments and analyze the results. Section V draws a final conclusion.

II. Related Work

A. Self-Supervised Graph Pre-Training

Self-supervised graph pre-training (SGP), whose goal is to learn representations from supervised signals derived from the graph data itself, has made significant progress recently. With the powerful learning capability of graph neural networks (GNNs) [20], [21], [22], [23], [24], advanced studies in this field have recently achieved great success in recommendation system [25], feature selection [26], graph clustering [27], [28], knowledge graph [29], [30], etc. One of the most representative self-supervised learning paradigms is contrastive SGP, where discriminative features are learned by pulling together the representations of semantically similar samples while pushing away the ones of unrelated samples [31], [32], [33], [34], [35], [36], [37], [38]. However, the impressive performance of these methods heavily relies on careful data augmentations, large amounts of negative samples, or relatively complicated

optimization strategies, which usually cause time- and resource-consuming issues. Comparatively, generative/predictive SGP methods [9], [39], [40], [41], [42], [43], [44] could naturally avoid the low-efficiency problem, as their optimization target is to reconstruct the input (or masked) information directly. In particular, masked graph autoencoders (MGAEs) [9], which aim to predict the masked content from the visible one, have significantly advanced classical graph autoencoders and shown their potential to achieve better performance against contrastive learning-based competitors.

B. Masked Graph Autoencoders

Masked signal modeling (MSM), which models masked signals locally to facilitate the extraction of significant features, has recently gained popularity in self-supervised vision and language applications, such as natural language understanding [5] and medical image analysis [4]. Inspired by the successes of existing masked autoencoders (MAEs) [1], [2], [3], [45], researchers pose a natural question regarding the potential of utilizing MAEs to handle large amounts of unlabelled graph data. To this end, MGAE [7] first applies an undirected edge-masking strategy to the original graph structure, and then utilizes a tailored cross-correlation decoder to predict the masked edges via a standard graph-based loss function. Similarly, MaskGAE [8] incorporates random corruption into the graph structure from both edge-wise level and path-wise level, and then utilizes edge-reconstruction and node-regression loss functions to match the predicted information with the original data. Another important research line in this field is node-masking-based MGAEs [9], [10], [46]. For example, GMAE [46] utilizes a graph transformer-based backbone [47] to learn representations and applies a cross-entropy loss to compare the reconstructed attributes with their ground truths. Similarly, GraphMAE [9] reconsiders the reconstruction loss functions of previous graph autoencoders and proposes an improved scaled cosine loss function to boost the quality of the masked attribute recovery. Towards this research line, some studies first randomly mask a portion of node attributes and connections of given graphs simultaneously, and then learn to predict the removed content from available information via two specific reconstruction objectives [12], [13]. More recently, MAEs-based techniques have inspired a broad range of graph learning applications. For instance, AutoCF [48] proposes an adaptive self-supervised augmentation for masked graph autoencoder pre-training in a recommendation system, achieving promising performance improvements while avoiding noise perturbation within raw data. GMAE-NAS [16] optimizes a masked graph autoencoder-enhanced predictor for neural architecture search by calculating the cross-entropy loss between the masked vertices of the original and reconstructed graphs. However, most of the aforementioned methods typically learn representations by minimizing the reconstruction loss in the raw data space directly, which may mislead the model into a local structural ambiguity situation caused by the non-Euclidean property of graphs. In contrast, RARE can effectively enhance the certainty in inferring masked data and the reliability of the self-supervision mechanism by reconciling the reconstructions of masked raw attributes and latent features, unleashing the potential of MAEs for graph analysis while mitigating the aforementioned negative effects.

C. Self-Distillation on Graphs

Self-distillation has emerged as a powerful technique for self-supervised learning, as evidenced by its successful applications in various domains [49], [50], [51], [52], [53]. This technique involves using the outputs of a target network as pseudo labels to guide the representation learning of an online network, which encourages the feature extractor to learn more generalized features. Self-distillation has also been widely developed and employed in multiple graph machine learning tasks, including graph structure learning [53] and augmentation-free node clustering [54]. While previous studies have focused on complete graphs, it is crucial to explore the effectiveness of self-distillation in boosting MGAEs in scenarios where the graph data is incomplete. This motivates us to investigate the feasibility of enhancing the robustness of MGAEs with the aid of self-distillation learning, as well as uncovering the key factors that contribute to the success of RARE.

III. Method

A. Task Definition and Overall Framework

1) Task Definition: In this study, we mainly focus on the task of self-supervised masked graph pre-training for unlabeled graphs. Our model is designed to learn two graph encoding functions (i.e., $\mathcal{F}_g(\cdot)$ and $\mathcal{F}_m(\cdot)$), along with a hidden predicting function (i.e., $\mathcal{F}_p(\cdot)$) to recover masked latent features from observations. Subsequently, a decoding function (i.e., $\mathcal{F}_d(\cdot)$) is employed to reconstruct the raw attributes of masked samples based on the predicted representations. The learned graph embedding can be saved and utilized for various downstream tasks, such as node classification and graph classification.

2) Overall Framework: As shown in Fig. 2, the learning procedure of RARE could be mainly grouped into three parts. The goal of the data masking part is to generate two complementary masked graphs by randomly masking some nodes with tokens under a mask ratio r. The masked latent feature completion part is the core of RARE, which aims to enhance the reliability of the self-supervision mechanism by leveraging more informative high-order sample correlations to drive the model learning. It consists of three components. First, the graph encoder maps the visible nodes into node representations. Second, the latent feature predictor performs a latent feature prediction from visible nodes to masked ones. Third, the momentum graph encoder receives the raw data of masked nodes as inputs and takes the resultant representations as implicit self-supervision signals for matching with the predicted representations. In the data decoding part, the decoder only maps the representations of masked nodes into the raw data space. Finally, $\mathcal{L}_M$ and $\mathcal{L}_R$ are minimized in the latent feature and raw data spaces, respectively. After pre-training, only the backbone of the graph encoder is adopted for downstream evaluations. The following subsections present the technical details of the corresponding components.

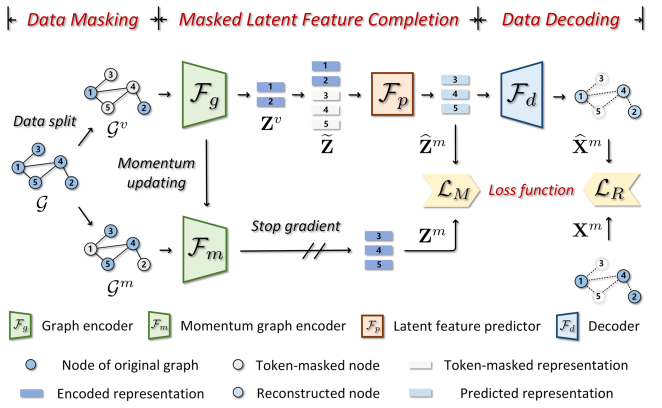

Fig. 2. Overview of RARE. During the pre-training phase, a graph is partitioned into two complementary masked graphs randomly. These graphs are then fed into the graph encoder and the momentum graph encoder, respectively. Next, the latent feature predictor is applied to predict the masked content from the visible one. Thereafter, the predicted representations are approximated to the output of the momentum graph encoder and processed by a simple decoder that reconstructs the raw attributes of masked nodes.

TABLE I
SUMMARY OF FREQUENTLY USED NOTATIONS

Notation	Description
$\widetilde{\mathbf{A}} \in \mathbb{R}^{\lvert\mathcal{V}\rvert\times\lvert\mathcal{V}\rvert}$	Normalized adjacency matrix
$\mathbf{X} \in \mathbb{R}^{\lvert\mathcal{V}\rvert\times D}$	Raw attribute matrix of all nodes
$\mathbf{X}^v \in \mathbb{R}^{\lvert\mathcal{V}^v\rvert\times D}$	Raw attribute matrix of visible nodes
$\mathbf{X}^m \in \mathbb{R}^{\lvert\mathcal{V}^m\rvert\times D}$	Raw attribute matrix of masked nodes
$\widehat{\mathbf{X}}^m \in \mathbb{R}^{\lvert\mathcal{V}^m\rvert\times D}$	Reconstructed raw attribute matrix of masked nodes
$\mathbf{H}^m \in \mathbb{R}^{\lvert\mathcal{V}^m\rvert\times d}$	Token-masked representation matrix
$\mathbf{Z}^v \in \mathbb{R}^{\lvert\mathcal{V}^v\rvert\times d}$	Representation matrix of visible nodes
$\mathbf{Z}^m \in \mathbb{R}^{\lvert\mathcal{V}^m\rvert\times d}$	Representation matrix of masked nodes
$\widetilde{\mathbf{Z}} \in \mathbb{R}^{\lvert\mathcal{V}\rvert\times d}$	Recomposed representation matrix
$\widehat{\mathbf{Z}}^m \in \mathbb{R}^{\lvert\mathcal{V}^m\rvert\times d}$	Predicted representation matrix of masked nodes
$\mathbf{b} \in \mathbb{R}^{\lvert\mathcal{V}\rvert}$	Random binary mask vector
$\mathbf{x}_i^v \in \mathbb{R}^D$	Raw attribute vector of ith visible node
$\mathbf{x}_j^m \in \mathbb{R}^D$	Raw attribute vector of jth masked node
$\widehat{\mathbf{x}}_j^m \in \mathbb{R}^D$	Reconstructed raw attribute vector of jth masked node
$\mathbf{h}_j^m \in \mathbb{R}^d$	Token-masked representation vector of jth masked node
$\mathbf{z}_j^m \in \mathbb{R}^d$	Representation vector of jth masked node
$\widehat{\mathbf{z}}_j^m \in \mathbb{R}^d$	Predicted representation vector of jth masked node

B. Data Masking

Before pre-training RARE, we first generate two complementary masked graphs as inputs. These graphs are then fed into the graph encoder $\mathcal{F}_g(\cdot)$ and the momentum graph encoder $\mathcal{F}_m(\cdot)$, respectively. To begin, we denote an undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ that contains $|\mathcal{V}|$ nodes with C categories. Here, $\mathcal{V}$ and $\mathcal{E}$ indicate the node set and the edge set, respectively. Generally, $\mathcal{G}$ can be characterized by its normalized adjacency matrix $\widetilde{\mathbf{A}} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ and raw attribute matrix $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}|\times D}$, where D refers to the dimension of node attributes. To perform the data-masking operation on a given graph, we initially draw a random binary mask vector $\mathbf{b} \in \mathbb{R}^{|\mathcal{V}|}$, where $b_i = 0$ if $\mathbf{x}_i$ is masked with a node token, and $b_i = 1$ otherwise. The probability of drawing 0 is r, which represents the mask ratio. Based on $\mathbf{b}$, we obtain the raw attribute matrix of visible nodes $\mathbf{X}^v \in \mathbb{R}^{|\mathcal{V}^v|\times D}$ and the raw attribute matrix of masked nodes $\mathbf{X}^m \in \mathbb{R}^{|\mathcal{V}^m|\times D}$

$$\mathbf{X}^v = \mathbf{X}[\mathbf{b}], \quad \mathbf{X}^m = \mathbf{X}[1-\mathbf{b}]. \tag{1}$$

Accordingly, the nodes in $\mathcal{G}$ are randomly divided into two sets, i.e., the visible node set $\mathcal{V}^v = \{\mathbf{x}_i^v\}_{i=1}^{|\mathcal{V}^v|}$ and the masked node set $\mathcal{V}^m = \{\mathbf{x}_j^m\}_{j=1}^{|\mathcal{V}^m|}$, where $\mathcal{V} = \mathcal{V}^v \cup \mathcal{V}^m$, $\mathcal{V}^v \cap \mathcal{V}^m = \varnothing$. $|\mathcal{V}^v|$ and $|\mathcal{V}^m|$ denotes the number of visible nodes and masked nodes, respectively, where $|\mathcal{V}| = |\mathcal{V}^v| + |\mathcal{V}^m|$. Correspondingly, we define token-masked node sets of the visible part and the masked part as $\mathcal{T}^v = \{\mathbf{t}_i^v\}_{i=1}^{|\mathcal{V}^v|}$ and $\mathcal{T}^m = \{\mathbf{t}_j^m\}_{j=1}^{|\mathcal{V}^m|}$, respectively, where $\mathbf{t}_i^v$ (or $\mathbf{t}_j^m$) $\in \mathbb{R}^D$ refers to a stochastic learnable vector. With these mathematical formulations, two complementary masked graphs used for pre-training can be denoted as $\mathcal{G}^v = \{\mathcal{V}^v, \mathcal{T}^m, \mathcal{E}\}$ and $\mathcal{G}^m = \{\mathcal{V}^m, \mathcal{T}^v, \mathcal{E}\}$, respectively. All frequently used notations are listed in Table I.

C. Masked Latent Feature Completion

As discussed in the previous section, although the masked autoencoder has proven effective and efficient, applying it to process non-Euclidean graphs directly may not always provide the required expressive capability for feature extraction. To address this issue, we propose a simple yet effective **M**asked **L**atent **F**eature **C**ompletion (MLFC) scheme. It facilitates model learning by incorporating more informative high-order sample correlations that are hard to be observed from the raw data perspective, leading to enhanced certainty in inferring masked content and a more reliable self-supervision mechanism for greater information encoding capability. The learning process of MLFC includes the following four main steps.

1) Graph Encoding: The graph encoder $\mathcal{F}_g(\cdot)$ is responsible for transforming the masked graph $\mathcal{G}^v$ into a low-dimension latent space. To achieve this, we employ a graph neural network (GNN)-based architecture that consists of a sequence of graph attention layers [55] or graph isomorphism layers [56] as the encoder backbone. Inspired by BYOL [51], we incorporate a multilayer perception (MLP) layer as a projector following the backbone to form the graph encoder. This encoder generates the representation matrix of visible nodes $\mathbf{Z}^v \in \mathbb{R}^{|\mathcal{V}^v|\times d}$, where d represents the latent dimension.

2) Latent Feature Predicting: Following the graph encoder $\mathcal{F}_g(\cdot)$, an autoencoder-style latent feature predictor $\mathcal{F}_p(\cdot)$ is elaborately designed, which consists of two parts, i.e., a graph attention (or graph isomorphism) layer that recovers the masked content from observations and an MLP layer that predicts the latent features of masked nodes based on recovered information. Specifically, we utilize a Concat function $C(\cdot)$ to integrate $\mathbf{Z}^v$ and a token-masked representation matrix $\mathbf{H}^m \in \mathbb{R}^{|\mathcal{V}^m|\times d}$, where $\mathbf{h}_j^m \in \mathbb{R}^d$ denotes a d-dimensional stochastic learnable vector. It is worth noting that the information concatenation used here to construct the recomposed representation matrix $\widetilde{\mathbf{Z}} \in \mathbb{R}^{|\mathcal{V}|\times d}$ is not the classic channel-wise or row-wise concatenation. Instead, we fill the visible part with $\mathbf{Z}^v$ and the masked part with $\mathbf{H}^m$ to create $\widetilde{\mathbf{Z}}$. Finally, we apply $\mathcal{F}_p(\cdot)$ to process $\widetilde{\mathbf{Z}}$ and obtain the predicted representation matrix of masked nodes $\widehat{\mathbf{Z}}^m \in \mathbb{R}^{|\mathcal{V}^m|\times d}$.

3) Momentum Graph Encoding: Since we have obtained the predicted representations of masked nodes, a natural question arises: how can we provide effective supervision to guide the masked latent feature completion in the unsupervised scenario? Our answer is to acquire the self-supervision signals from the data itself via self-distillation learning. To this end, we introduce a momentum graph encoder $\mathcal{F}_m(\cdot)$ that has the same architecture as the graph encoder. This encoder is responsible for encoding the raw data of masked samples and utilizing their complete representations to provide the predicted ones with stable optimization guidance. Concretely, we take the masked graph $\mathcal{G}^m$ as an input and feed it into $\mathcal{F}_m(\cdot)$. The resultant representation matrix $\mathbf{Z}^m \in \mathbb{R}^{|\mathcal{V}^m| \times d}$ preserves high-order sample correlations and serves as implicit self-supervision signals to refine $\widehat{\mathbf{Z}}^m$. It is worth noting that $\mathcal{F}_m(\cdot)$ is detached from the gradient back-propagation, and its parameters are updated by exponential moving average (EMA) [51]. The parameters of $\mathcal{F}_g(\cdot)$ and $\mathcal{F}_m(\cdot)$ are denoted as $\Theta_{\mathcal{F}_g}$ and $\Theta_{\mathcal{F}_m}$, respectively, and the parameters of $\mathcal{F}_m(\cdot)$ are updated by

$$\Theta_{\mathcal{F}_m} \leftarrow \mu\Theta_{\mathcal{F}_m} + (1-\mu)\Theta_{\mathcal{F}_g}, \tag{2}$$

where μ denotes the momentum factor that has been determined empirically and fixed as 0.1. Since both graph encoders involve training on multiple subsets of a common graph, the EMA can provide $\mathcal{F}_m(\cdot)$ with a smooth estimate of the underlying graph data distribution from $\mathcal{F}_g(\cdot)$, thus promoting $\mathcal{F}_m(\cdot)$ to encode reliable information.

4) Latent Feature Matching: This operation acts as an implicit form of self-supervision for recovering masked content at the feature level. Rather than aiming to maintain similarity between the reconstructed nodes (or edges) and the raw information of the original graph, our method focuses on ensuring that the predicted representations precisely match with the underlying structural statistics calculated by the momentum graph encoder. To this end, we approximate $\widehat{\mathbf{Z}}^m$ to $\mathbf{Z}^m$ by minimizing the following formulation:

$$\mathcal{L}_M = \frac{1}{|\mathcal{V}^m|}\sum_{j=1}^{|\mathcal{V}^m|} \|\widehat{\mathbf{z}}_j^m - \mathbf{z}_j^m\|^2, \tag{3}$$

where $\widehat{\mathbf{z}}_j^m \in \mathbb{R}^d$ and $\mathbf{z}_j^m \in \mathbb{R}^d$ indicate the representation vectors of jth masked node within $\widehat{\mathbf{Z}}^m$ and $\mathbf{Z}^m$, respectively. According to (3), we recover masked information within the latent space by constraining the predicted representations to match with the ones that preserve more informative underlying structural information of the graph. By taking these implicit self-supervision signals as model learning guidance, we could ensure the integrity and accuracy of predicted representations, resulting in a higher quality of the learned graph embedding.

D. Data Decoding

To complete the raw attributes of masked nodes, we employ a simple MLP layer as a decoder $\mathcal{F}_d(\cdot)$ to map $\widehat{\mathbf{Z}}^m$ into the raw data space. Once the reconstructed raw attribute matrix of masked nodes $\widehat{\mathbf{X}}^m \in \mathbb{R}^{|\mathcal{V}^m| \times D}$ has been processed by the decoder, we take $\mathbf{X}^m \in \mathbb{R}^{|\mathcal{V}^m| \times D}$ as explicit self-supervision signals and reconstruct the raw data for masked nodes by minimizing the distance between $\widehat{\mathbf{X}}^m$ and $\mathbf{X}^m$. Inspired by the SCE loss [9], we design an improved scaled cosine loss function to boost the stability of network training, formulated as

$$\mathcal{L}_R = -\frac{1}{|\mathcal{V}^m|}\sum_{j=1}^{|\mathcal{V}^m|} \log\left(\frac{1}{2} + \frac{\langle \widehat{\mathbf{x}}_j^m, \mathbf{x}_j^m\rangle}{2\|\widehat{\mathbf{x}}_j^m\|\|\mathbf{x}_j^m\|}\right)^t, \tag{4}$$

where $\langle\cdot,\cdot\rangle$ refers to an inner product operation. $\widehat{\mathbf{x}}_j^m \in \mathbb{R}^D$ and $\mathbf{x}_j^m \in \mathbb{R}^D$ denotes the raw attribute vectors of jth masked node within $\widehat{\mathbf{X}}^m$ and $\mathbf{X}^m$, respectively. t is a scaling factor, and we empirically set $t = 2$ in most cases.

Algorithm 1: Robust Masked Graph Autoencoder (RARE).

Input: Raw graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$; token-masked nodes $\{\mathcal{T}^v, \mathcal{T}^m\}$; token-masked representation matrix $\mathbf{H}^m$; maximum iterations E; mask ratio r; scaling factor t; balanced coefficient α; model parameters $\{\Theta_{\mathcal{F}_g}, \Theta_{\mathcal{F}_m}, \Theta_{\mathcal{F}_p}, \Theta_{\mathcal{F}_d}\}$; learning rate η.
Output: Pre-trained parameters $\Theta_{\mathcal{F}_g}$.

```
Initialize {Θ_{F_g}, Θ_{F_m}, Θ_{F_p}, Θ_{F_d}} with an Xavier manner.
for e = 1 to E do
    {V^v, V^m} ← Split V into visible and masked node sets.
    {G^v, G^m} ← Obtain two masked graphs.
    Z^v ← Obtain representations from G^v with F_g(·).
    Z̃ ← Integrate Z^v and H^m using C(·).
    Ẑ^m ← Obtain predicted representations with F_p(·).
    Z^m ← Obtain representations from G^m with F_m(·).
    L_M ← Calculate the loss by (3).
    X̂^m ← Obtain raw attributes from Ẑ^m with F_d(·).
    L_R ← Calculate the loss by (4).
    L ← Calculate the total loss by (5).
    Update {Θ_{F_g}, Θ_{F_p}, Θ_{F_d}} by calculating:
    Θ_{F_g} ← Θ_{F_g} − η∇_{Θ_{F_g}} L;
    Θ_{F_p} ← Θ_{F_p} − η∇_{Θ_{F_p}} L;
    Θ_{F_d} ← Θ_{F_d} − η∇_{Θ_{F_d}} L.
    Update Θ_{F_m} by (2).
end for
return Θ_{F_g}
```

E. Loss Function and Complexity Analysis

1) Loss Function: By integrating implicit and explicit self-supervision mechanisms in a united pre-training framework, the total loss of the proposed RARE can be formulated as a weighted combination of the latent feature matching loss $\mathcal{L}_M$ and the raw attribute reconstruction loss $\mathcal{L}_R$

$$\mathcal{L} = \mathcal{L}_M + \alpha\mathcal{L}_R, \tag{5}$$

where α is a balanced coefficient. In the inference phase, the input graph $\mathcal{G}$ with $\widetilde{\mathbf{A}}$ and $\mathbf{X}$ is fed into RARE without any

data-masking operations. The resultant graph embedding can be saved and used for downstream evaluations, such as graph classification and image recognition tasks. The detailed pre-training procedure and pseudo code of RARE are illustrated in Algorithm 1 and Appendix A, available online, respectively.

2) Complexity Analysis: The time complexity of the proposed RARE could be discussed from the following two perspectives: the graph autoencoder framework and the loss function computation. For two graph encoders, the complexities of $\mathcal{F}_g(\cdot)$ and $\mathcal{F}_m(\cdot)$ are $\mathcal{O}((|\mathcal{V}|d^2+|\mathcal{E}|d)KL_e)$, where $|\mathcal{V}|$, $|\mathcal{E}|$, L_e, and K are the number of nodes, edges, encoder layers, and attention heads, respectively. d is the dimensions of sample features. For the latent feature predictor, the complexity of $\mathcal{F}_p(\cdot)$ is $\mathcal{O}((|\mathcal{V}|d^2+|\mathcal{E}|d)K+|\mathcal{V}|d^2)$. For the decoder, the complexity of $\mathcal{F}_d(\cdot)$ is $\mathcal{O}(|\mathcal{V}|d^2L_d)$, where L_d is the number of decoder layers. For the computation of the loss function, the time complexities of $\mathcal{L}_M$ and $\mathcal{L}_R$ are $\mathcal{O}(|\mathcal{V}^m|d)$, where $|\mathcal{V}^m|$ is the number of masked nodes. Therefore, the overall time complexity of RARE for each training iteration is $\mathcal{O}((|\mathcal{V}|d^2+|\mathcal{E}|d)KL_e+|\mathcal{V}|d^2L_d+|\mathcal{V}^m|d)$. We can observe that the complexity of RARE is linear with both the number of nodes $|\mathcal{V}|$ and edges $|\mathcal{E}|$ of the graph, making the proposed RARE theoretically efficient and scalable. For more time complexity discussions among different masked graph autoencoders, please refer to Appendix B, available online.

F. Discussion

In this section, we aim to explain the reasons why the proposed MLFC scheme is effective and why RARE works better than existing MGAEs, respectively.

1) Why The Proposed MLFC Scheme Is Effective: We start from a more intuitive masked signal modeling perspective to revisit MLFC. As aforementioned, we can obtain the latent features of visible and masked nodes from two complementary masked views (i.e., $\mathcal{G}^v$ and $\mathcal{G}^m$), respectively

$$\mathbf{Z}^v=\mathcal{F}_g(\mathbf{X}^v,\mathbf{T}^m,\widetilde{\mathbf{A}}), \tag{6}$$

$$\mathbf{Z}^m=\mathcal{F}_m(\mathbf{X}^m,\mathbf{T}^v,\widetilde{\mathbf{A}}). \tag{7}$$

After that, $\mathcal{F}_p(\cdot)$ outputs the predicted representation matrix of masked nodes $\widehat{\mathbf{Z}}^m$ from that of visible ones $\mathbf{Z}^v$, and then match $\widehat{\mathbf{Z}}^m$ with $\mathbf{Z}^m$ by (3), which can be rewritten as

$$\mathcal{L}_M=\mathbb{E}_{\mathbf{Z}^v,\mathbf{Z}^m}\|\mathcal{F}_p(\mathbf{Z}^v,\mathbf{H}^m,\widetilde{\mathbf{A}})-\mathbf{Z}^m\|^2. \tag{8}$$

Based on (8), we observe that the MLFC scheme actually learns to pair two complementary views (i.e., $\mathbf{Z}^v$ and $\mathbf{Z}^m$) through a latent feature matching task. Inspired by the previous work [57], [58], we denote a bipartite graph $\mathcal{G}_B=\{\mathcal{Z}^v,\mathcal{Z}^m,\mathcal{E}_B\}$ to model the corresponding learning problem, where $\mathcal{Z}^v=\{\mathbf{Z}^v\}_{i=1}^{|\mathcal{Z}^v|}$ and $\mathcal{Z}^m=\{\mathbf{Z}^m\}_{j=1}^{|\mathcal{Z}^m|}$ denotes the sets of visible views and masked views, respectively. $\mathcal{E}_B$ is formulated as an adjacency matrix $\mathbf{A}_B\in\mathbb{R}^{|\mathcal{Z}^m|\times|\mathcal{Z}^v|}$, whose normalized version is $\widetilde{\mathbf{A}}_B=\mathbf{D}^{m-\frac{1}{2}}\mathbf{A}_B\mathbf{D}^{v-\frac{1}{2}}$. Here, both $\mathbf{D}^m$ and $\mathbf{D}^v$ are diagonal degree matrices of $\mathbf{A}_B$. Consequently, we can derive an asymmetric instance alignment loss between $\mathbf{Z}^v$ and $\mathbf{Z}^m$ to lower bounded $\mathcal{L}_M$.

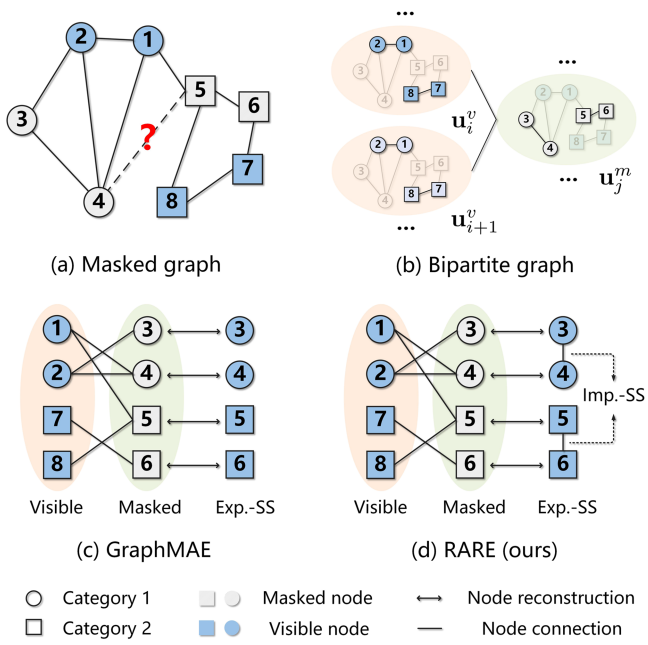

Fig. 3. Motivation illustration: (a) A masked graph; (b) A bipartite graph that includes two types of complementary views; (c) and (d) Comparison of self-supervision mechanisms between two methods. To drive the model learning, GraphMAE [9] only conducts an explicit self-supervision (i.e., Exp.-SS) mechanism by matching the predicted nodes with raw ones, while the proposed RARE differs from GraphMAE in taking high-order sample correlations as implicit self-supervision (i.e., Imp.-SS) signals that are hard to be observed from the raw data perspective.

Theorem 1: We assume that any autoencoder-style architecture $\mathcal{F}(\cdot)$ satisfies $\mathbb{E}_{\mathbf{X}}\|\mathcal{F}(\mathbf{X})-\mathbf{X}\|^2\le\delta$, where $\mathbf{X}$ represents either visible content $\mathbf{Z}^v$ or masked content $\mathbf{Z}^m$, therefore, the MLFC loss on the bipartite graph $\mathcal{G}_B$ can be lower bounded by

$$\begin{aligned}\mathcal{L}_M&\ge-\mathbb{E}_{\mathbf{Z}^v,\mathbf{Z}^m}\mathcal{F}_p(\mathbf{Z}^v,\mathbf{H}^m,\widetilde{\mathbf{A}})^\top\mathcal{F}(\mathbf{Z}^m)-\delta+1\\&\ge-\mathrm{tr}(\mathbf{U}^{m\top}\widetilde{\mathbf{A}}_B\mathbf{U}^v)-\delta+1,\end{aligned} \tag{9}$$

where $\mathbf{U}^v\in\mathbb{R}^{|\mathcal{Z}^v|\times(|\mathcal{V}^m|d)}$ denotes the output matrix of $\mathcal{F}_p(\cdot)$ on $\mathcal{Z}^v$ whose i-row is $\mathbf{u}_i^v=\sqrt{d_i}\mathcal{F}_p(\mathbf{Z}^v,\mathbf{H}^m,\widetilde{\mathbf{A}})_i\in\mathbb{R}^{|\mathcal{V}^m|d}$, and $\mathbf{U}^m\in\mathbb{R}^{|\mathcal{Z}^m|\times(|\mathcal{V}^m|d)}$ denotes the output matrix of $\mathcal{F}(\cdot)$ on $\mathcal{Z}^m$ whose j-row is $\mathbf{u}_j^m=\sqrt{d_j}\mathcal{F}(\mathbf{Z}^m)_j\in\mathbb{R}^{|\mathcal{V}^m|d}$. Note that both $\mathcal{F}_p(\cdot)$ and $\mathcal{F}(\cdot)$ are normalized.

According to (9), it is evident that the MLFC scheme aims to minimize $\mathcal{L}_M$ by conducting an implicit similarity-based alignment between connected visible and masked views in the latent space through an autoencoder-style predictor. Intuitively, as illustrated in Fig. 3(b), we consider a pair of second-order neighbor visible views (e.g., $\mathbf{u}_i^v$ and $\mathbf{u}_{i+1}^v$) that share a common complementary masked view (e.g., $\mathbf{u}_j^m$). By enforcing $\mathbf{u}_i^v$ and $\mathbf{u}_{i+1}^v$ to recover $\mathbf{u}_j^m$ simultaneously, the latent features of visible views are implicitly correlated through the MLFC scheme. In other words, the second-order neighbors act as positive sample pairs that are pulled closer as the instance alignment (i.e., positive samples should remain close in the latent space) in non-contrastive learning [51], [52]. From this perspective,

the proposed MLFC scheme serves as a hidden regularization that helps the model encode useful features, thereby promoting greater information encoding capability for better downstream performance. Please see Appendix C, available online, for proof details of Theorem 1 and more theoretical discussions.

2) Why RARE Works Better Than Existing MGAEs: To better understand the superiority of RARE compared to its competitors, we conduct a comparison of the self-supervision mechanism between the proposed RARE and the state-of-the-art (SOTA) GraphMAE [9] via a toy example illustrated in Fig. 3. As previously mentioned, completing masked content within a masked graph can be regarded as a complementary view pairing problem between the visible and masked data. To solve this problem, GraphMAE [9] adopts an explicit self-supervision (i.e., Exp.-SS) mechanism by matching predicted nodes with raw ones directly, which may mislead the model into a local structural ambiguity situation. For example, as shown in Fig. 3(c), two masked samples (e.g., Node 4 and Node 5) belonging to different categories share a common visible sample (e.g., Node 1). By enforcing Node 1 to reconstruct Node 4 and Node 5, GraphMAE subconsciously reduces their distance in the latent space. Consequently, the model may struggle to differentiate Node 4 and Node 5 accurately in the unsupervised scenario. This is because the self-supervision signals provided by GraphMAE only preserve raw data information that is insufficient to ascertain whether two nodes belong to the same category or not. In contrast, RARE employs both implicit and explicit self-supervision mechanisms for masked content recovery by performing a joint mask-then-reconstruct strategy in both latent feature and raw data spaces. Particularly, in RARE, the MLFC scheme could take more informative high-order sample correlations as implicit self-supervision signals, which are not readily available in the raw data space. As shown in Fig. 3(d), by incorporating prior knowledge that 1) Node 3 or Node 6 is closely related to Node 4 or Node 5; and 2) Node 3 and Node 6 are highly likely not to belong to the same category, it would become easier for the model to infer the relationship between Node 4 and Node 5. As a result, the two nodes are pushed apart from each other in the latent space. Extensive empirical results support our claim that the proposed RARE indeed works better than GraphMAE by making the learned samples belonging to different categories more distinguishable.

IV. Experiments

In this section, we evaluate the effectiveness of RARE against advanced SGP methods. In the following, we begin with a brief introduction to experimental setups, including benchmark datasets, implementation procedures, training setups, and baseline methods. Then, we report experimental results with corresponding analyses.

A. Evaluation Setups

1) Benchmark Datasets: We conduct experiments to compare the proposed RARE with several baseline methods on seventeen datasets in total, including seven node classification datasets, seven graph classification datasets, and three image classification datasets. Detailed data statistics and affinity graph construction for non-graph datasets are presented in Appendix D, available online.

- *Citation Graphs:* In citation graphs, nodes typically represent papers, where node attributes correspond to keywords extracted from the papers. Furthermore, edges denote cross-citation connections, while categories reflect the topics covered in the papers. Please note that nodes within citation graphs may occasionally represent authors, institutions, or other entities [59]. The used citation graphs include Cora, Citeseer, Pubmed, Wiki-CS, and Corafull.
- *Social Graphs:* The social graph represents entities (e.g., users) as nodes, where their interests are captured as attributes, and their social interactions are represented as edges [59]. The used social graphs include Flickr, Yelp, IMDB-B, IMDB-M, and COLLAB.
- *Molecule Graphs:* In molecular graphs, nodes represent individual atoms within the molecule, and the atom index is denoted as the node attribute. The edges in the graph correspond to the bonds between the atoms. Molecular graphs usually consist of multiple inter-connected graphs [59]. The used molecule graphs include MUTAG, PTC-MR, and NCI1.
- *Protein Graphs:* The protein molecule graph is a specialized form of a molecule graph, where nodes represent amino acids, and an edge signifies that the connected nodes are within a distance of less than six angstroms [59]. The used protein network dataset is PROTEINS.
- *Image datasets:* Usps comprises 9,298 Gy-scale handwritten digit images, each with dimensions of 16×16 pixels. The features represent the gray values of the pixel points in the images [60]. Mnist, which is a subset of a larger collection from NIST, consists of 60,000 training samples and 10,000 test samples of handwritten digits. The digits in Mnist have been resized, normalized, and centered within fixed-size images of 28×28 pixels [61]. Fashion-mnist includes 60,000 training images and 10,000 test images of fashion and clothing items. Each image is standardized to a size of 28×28 pixels in grayscale. Fashion-mnist was developed by Zalando as a compatible replacement for the original Mnist [62].

2) Implementation Procedures: The learning procedure of RARE mainly includes two steps: 1) in the pre-training task, all samples of datasets except for Flickr and Yelp are fed into the proposed RARE for at least 20 training iterations by minimizing (5). Since Flickr and Yelp are commonly used for inductive evaluations, we follow the public data split as GraphSAINT [63], where 50%/75%, 25%/10%, and 25%/15% nodes are randomly sampled to form the train, validation, and test sets on Flickr and Yelp, respectively; and 2) in the node classification and image classification tasks, we use the publicly available train/validation/test data split for Cora, Citeseer, Pubmed, Flickr, and Yelp. For WikiCS, Corafull, Usps, Mnist, and Fashion-mnist, since these datasets have no publicly available data split, we perform a random data split where 7%, 7%, and 86% nodes are randomly sampled to form the train, validation, and test sets, respectively. We train a simple linear

classifier with Adam optimizer until convergence by optimizing a cross-entropy loss by 10 times. For graph classification, all used datasets are partitioned based on 10-fold cross-validation for training and testing. As is done in GraphMAE [9], we take the support vector machine (SVM) as a classifier and record the results with 10-fold cross-validation after 5 separate runs. To mitigate the adverse impact of randomness, we report average accuracy (ACC) values with standard deviations for each model in all downstream evaluations.

3) Training Setups: To ensure a fair comparison, all experiments are conducted on the same device. For all compared baselines, we directly report the results listed in the existing literature if available. Otherwise, we implement their official source codes and report the reproduced performance. For our method, we perform a grid search to select hyper-parameters on the following searching space: the mask ratio r is selected between {0.5, 0.75}; the balanced coefficient α is searched from 1 to 9; the scale factor t is selected between {1, 2}; the hidden size of latent features is selected from {256, 512, 1024}; the number of feature extractor layers is selected from {1, 2, 3, 4, 5}; the momentum rate is empirically fixed to 0.1 by default; the learning rate of the Adam optimizer is selected from {1.5e-4, 5e-4, 1e-3}; the maximum epoch is determined according to the cases of model convergence. Particularly, in the graph classification task, we 1) choose the batch size from {32, 64}, similar to GraphMAE [9]; 2) consistently adopt a batch normalization operation to regularize the model learning; and 3) follow the sample pooling setups in GraphMAE, where a non-parameterized graph pooling function (e.g., max-pooling, mean-pooing or sum-pooling) is employed to generate graph-level representations. Please note that we employ similar hyper-parameter setups as reported in GraphMAE [9], and most hyper-parameters are not carefully tuned for the ease of model learning. More details can be found in Appendix E, available online, such as experimental infrastructures, result illustrations, and detailed hyper-parameter setups.

4) Baseline Methods: In our experiments, we compare the proposed RARE with twenty-three baseline methods. Detailed descriptions of all compared baseline methods are listed below.

- *GCN [64]* is a typical graph neural network (GNN) that utilizes the label information to train the network in an inductive manner.
- *GAT [55]* is a powerful graph neural network that utilizes attention mechanisms to capture the importance of neighboring nodes within a graph.
- *GAE [39]* utilizes a GCN-based graph encoder to extract useful information from the graph and then decodes it by an inner product operation.
- *DGI [31]* learns the graph embedding by achieving mutual information (MI) maximization between node representations and global summaries.
- *MVGRL [34]* makes the first attempts to introduce the concept of multi-view self-supervised graph pre-training (SGP) by discriminating the representations across two augmented graph views.
- *GRACE [65]* generates two augmented graph views at first and then designs an improved InfoNCE loss to maximize the agreement of their representations.
- *BGRL [52]* is an advanced SGP method without requiring negative samples, which gets rid of the potentially quadratic bottleneck.
- *InfoGCL [66]* reduces the redundant information between contrastive sample pairs while preserving as much task-relevant information as possible.
- *CCA-SSG [67]* is a simple, efficient, and effective SGP framework with a canonical correlation analysis (CCA)-based optimization target.
- *MaskGAE [8]* merely removes partial node connections and learns to predict the removed edges from observations via an edge-level reconstruction.
- *GraphMAE [9]* pre-trains a graph autoencoder to reconstruct masked attributes from visible ones in the raw data space.
- *SeeGera [13]* employs a variational inference framework to learn useful features via the structure/feature mask-then-reconstruct mechanism.
- *WL [68]* leverages a rapid feature extraction scheme based on the Weisfeiler-Lehman test of isomorphism on graphs.
- *DGK [69]* leverages the dependency information between sub-structures to extract useful sample features for graphs.
- *GIN [56]* is a simple neural graph isomorphism architecture, whose discriminative/representational power is equal to the power of the WL test.
- *DiffPool [70]* introduces a plug-and-play differentiable graph pooling module to extract the hierarchical structural information of the graph.
- *Graph2vec [71]* learns data-driven distributed representations of arbitrary-sized graphs in an unsupervised and task-agnostic manner.
- *Infograph [36]* introduces an MI estimation mechanism to encourage the sub-graph embedding to capture the global properties of the graph.
- *GraphCL [32]* learns node representations across two augmented graph views by maximizing the MI agreement between two-view representations.
- *JOAO [72]* is a unified bi-level optimization framework that automatically conducts graph augmentations to learn meaningful graph embedding.
- *GCC [33]* is a representative SGP method that captures the universal structural properties of graphs across multiple instances.
- *VGG16 [73]* is a classical deep convolutional network with small (3×3) convolution filters.
- *ResNet18 [74]* is a popular deep convolutional network with hundreds of layers, where skip connections or shortcuts are used to jump over some layers.

B. Overall Performance

1) Evaluation on Node Classification: As shown in Table II, we report the node classification performance of thirteen compared methods on seven datasets. From these results, several major observations can be concluded: 1) the proposed RARE consistently outperforms two supervised methods on all datasets, with margins going up to 7.7%–14.8% on Flickr and Yelp. These improvements demonstrate the great potential of

TABLE II
NODE CLASSIFICATION PERFORMANCE COMPARISON

Learning Type	Method	Cora	Citeseer	Pubmed	WikiCS	Corafull	Flickr	Yelp
Supervised	GCN [65]	81.5	70.3	79.0	66.7±0.5	48.6±0.5	42.9±0.1	57.3±0.1
	GAT [56]	83.0±0.7	72.5±0.7	79.0±0.3	69.4±1.0	50.7±0.2	43.9±0.1	57.6±0.1
Self-supervised	GAE [39]	71.5±0.4	65.8±0.4	72.1±0.5	67.3±0.3	52.0±0.1	-	-
	DGI [31]	82.3±0.6	71.8±0.7	76.8±0.6	64.8±0.6	48.2±0.5	45.0±0.2	57.4±0.1
	MVGRL [34]	83.5±0.4	73.3±0.5	80.1±0.7	64.8±0.7	52.6±0.5	-	-
	GRACE [66]	81.9±0.4	71.2±0.5	80.6±0.4	68.0±0.7	45.2±0.1	-	-
	BGRL [53]	82.7±0.6	71.1±0.8	79.6±0.5	65.5±1.5	47.4±0.5	39.4±0.1	-
	InfoGCL [67]	83.5±0.3	73.5±0.4	79.1±0.2	-	-	-	-
	CCA-SSG [68]	84.0±0.4	73.1±0.3	81.0±0.4	67.4±0.9	53.5±0.4	49.1±0.1	59.6±0.1
	SeeGera [13]	82.8±0.3	71.6±0.2	79.2±0.3	65.8±0.2	-	-	-
	MaskGAE [8]	82.6±0.3	73.4±0.6	81.0±0.3	66.0±0.2	52.2±0.1	49.1±0.1	68.1±0.1
	GraphMAE [9]	**84.2±0.4**	73.4±0.4	81.1±0.4	65.7±0.7	53.4±0.1	49.6±0.2	69.4±0.2
	Ours	**84.2±0.2**	**74.1±0.3**	**81.8±0.2**	**69.0±0.6**	**55.5±0.1**	**50.6±0.1**	**72.1±0.6**

"-" means unavailable source code or out-of-memory error. The boldface and underline values indicate the best and the runner-up results (%) of masked graph autoencoder methods, respectively.

TABLE III
GRAPH CLASSIFICATION PERFORMANCE COMPARISON

Learning Type	Method	IMDB-B	IMDB-M	PROTEINS	COLLAB	MUTAG	PTC-MR	NCI1
Supervised	GIN [57]	75.1±5.1	52.3±2.8	76.2±2.8	80.2±1.9	89.4±5.6	63.7±8.2	82.7±1.7
	DiffPool [69]	72.6±3.9	-	75.1±3.5	78.9±2.3	85.0±10.3	-	-
Graph Kernels	WL [70]	72.3±3.4	47.0±0.5	72.9±0.6	-	80.7±3.0	58.0±0.5	80.3±0.5
	DGK [71]	67.0±0.6	44.6±0.5	73.3±0.8	-	87.4±2.7	60.1±2.6	80.3±0.5
Self-supervised	Graph2vec [72]	71.1±0.5	50.4±0.9	73.3±2.1	-	83.2±9.3	60.2±6.9	73.2±1.8
	Infograph [36]	73.0±0.9	49.7±0.5	74.4±0.3	70.7±1.1	89.0±1.1	61.7±1.4	76.2±1.1
	GraphCL [32]	71.1±0.4	48.6±0.7	74.4±0.5	71.4±1.2	86.8±1.3	61.3±2.1	77.9±0.4
	JOAO [73]	70.2±3.1	49.2±0.8	74.6±0.4	69.5±0.3	87.4±1.0	-	78.1±0.5
	GCC [33]	72.0	49.4	-	78.9	-	-	-
	MVGRL [34]	74.2±0.7	51.2±0.5	-	-	89.7±1.1	62.5±1.7	-
	InfoGCL [67]	75.1±0.9	51.4±0.8	-	80.0±1.3	91.2±1.3	63.5±1.5	80.2±0.6
	GraphMAE [9]	75.5±0.7	51.6±0.5	75.3±0.4	80.3±0.5	88.2±1.3	57.6±0.8	80.4±0.3
	Ours	**76.2±0.1**	**53.1±0.1**	**76.4±0.2**	**81.2±0.6**	**88.6±0.6**	**59.3±0.7**	**81.3±0.4**

"-" means unavailable source code or out-of-memory error. The boldface and underline values indicate the best and the runner-up results (%) of masked graph autoencoder methods, respectively.

masked graph autoencoders for effectively handling massive unlabeled graph data; 2) InfoGCL is one of the strongest contrastive self-supervised graph pre-training methods, while the proposed RARE improves it by 0.7%, 0.6%, 2.7% accuracy on Cora, Citeseer, and Pubmed, respectively. This phenomenon indicates that RARE can effectively boost the learned representations by conducting a mask-then-reconstruct mechanism instead of relying on a relatively complicated contrastive learning mechanism; 3) taking the results on WikiCS for example, RARE significantly outperforms SeeGera, MaskGAE, and GraphMAE by 3.2%, 3.0%, and 3.3%, respectively. These benefits are attributed to the novel idea of integrating implicit and explicit self-supervision mechanisms to drive model learning by performing a joint mask-then-reconstruct strategy in both latent feature and raw data spaces; and 4) on Cora, RARE achieves competitive results compared to the most powerful masked graph autoencoder, i.e., GraphMAE. However, it is possible that the full potential of model optimization was not demonstrated due to the limited size of the test dataset. Expanding the size of the test dataset, for instance, by incorporating larger datasets like WikiCS and Yelp, could reveal that RARE has the potential to yield better performance.

2) Evaluation on Graph Classification: Table III summarizes graph classification results of thirteen methods on seven datasets. The results reveal several key observations that are similar to those obtained from the node classification task: 1) the performance of RARE is highly competitive compared to both graph kernel-based methods and supervised methods, indicating that the masked graph autoencoder has the potential to be a promising alternative for self-supervised graph pre-training; 2) compared to GraphCL and JOAO, our method achieves significant performance improvements (up to 1.8%–11.7%) over them on almost all datasets. However, some contrastive learning methods, such as InfoGCL and MVGRL, demonstrate better performance than masked graph autoencoders on MUTAG and PTC-MR. This may be because in some cases, the partitioning of small-scale graph data can be easily achieved through multi-view contrastive learning; and 3) RARE achieves an approximate 1.1% average performance gain over GraphMAE on all datasets, which further indicates that RARE can effectively leverage both implicit and explicit self-supervision signals to improve the quality of the learned graph embedding.

3) Evaluation on Image Classification: To verify the superiority of RARE in-depth, Table IV reports the image classification performance of six methods on three datasets. From those results, we can obtain the following observations: 1) the proposed RARE shows a significant advantage against existing state-of-the-art MGAEs and other baselines on all image

TABLE IV
IMAGE CLASSIFICATION PERFORMANCE COMPARISON

Learning Type	Method	Usps	Mnist	Fashion-mnist
Supervised	VGG16 [74]	94.5±0.5	95.6±0.8	85.6±0.6
	ResNet18 [75]	94.2±1.6	96.1±1.3	85.4 ±1.5
Self-supervised	GAE [39]	75.8±0.2	-	-
	MaskGAE [8]	92.2±0.2	87.1±0.1	78.5±0.1
	GraphMAE [9]	93.0±0.5	91.7±0.1	79.8±0.2
	Ours	**94.3±0.3**	**94.2±0.2**	**85.7±0.1**

"-" means the out-of-memory error. The boldface and underline values indicate the best and the runner-up results (%) of masked graph autoencoder methods, respectively.

TABLE V
ABLATION STUDY FOR THE MLFC SCHEME

Dataset	w/o-Pred.	w/o-Mome.	w-$\mathcal{F}_m(\mathcal{G})$	Ours
Cora	82.5 (1.7 ↓)	83.4 (0.8 ↓)	83.7 (0.5 ↓)	**84.2**
Citeseer	72.7 (1.4 ↓)	70.5 (3.6 ↓)	73.6 (0.5 ↓)	**74.1**
Pubmed	79.5 (2.3 ↓)	78.8 (3.0 ↓)	81.4 (0.4 ↓)	**81.8**
Corafull	52.2 (3.3 ↓)	51.9 (3.6 ↓)	55.2 (0.3 ↓)	**55.5**
IMDB-B	74.6 (1.6 ↓)	74.7 (1.5 ↓)	75.6 (0.6 ↓)	**76.2**
IMDB-M	51.7 (1.4 ↓)	50.8 (2.3 ↓)	52.3 (0.8 ↓)	**53.1**
PROTEINS	74.1 (2.3 ↓)	75.6 (0.8 ↓)	75.4 (1.0 ↓)	**76.4**
PTC-MR	54.1 (5.2 ↓)	57.5 (1.8 ↓)	58.0 (1.3 ↓)	**59.3**

"w / o-Pred." and "w / o-Mome." denote two RARE variants with the latent feature predictor and the momentum graph encoder being removed, respectively."w-$\mathcal{F}_m(\mathcal{G})$" is a variant of RARE whose momentum graph encoder accepts a complete graph $\mathcal{G}$.↓ denotes the performance degradation. The boldface values indicate the best results (%).

benchmarks. For example, on Mnist and Fashion-mnist datasets, RARE consistently outperforms the best edge-masking-based MaskGAE and node-masking-based GraphMAE by 7.1%/2.5% and 7.2%/5.9% accuracy, respectively. These improvements once again demonstrate the effectiveness of introducing implicit self-supervision signals for model learning; and 2) it is interesting to note that RARE can achieve competitive or slightly better results than typical supervised classification methods, such as VGG16 and ResNet18. This implies that improving the reliability of the self-supervision mechanism can facilitate RARE to unleash its potential for SGP. Thus, the learned representations show good robustness and generalization across a wide range of downstream tasks.

C. Ablation Study

1) Impact of The MLFC Scheme: To demonstrate the effectiveness of the proposed masked latent feature completion scheme, we compare RARE with its three variants on eight datasets. Concretely, "w/o-Pred." implies that RARE removes the latent feature predictor. "w/o-Mome." indicates that RARE discards the momentum graph encoder. "w-$\mathcal{F}_m(\mathcal{G})$" denotes that the momentum graph encoder of RARE accepts $\mathcal{G}$ rather than $\mathcal{G}^m$. As shown in Table V, some major observations can be summarized: 1) when compared to "w/o-Pred.", the latent feature predictor produces a performance gain of 1.4%–5.2% on eight datasets, indicating that this component plays a vital role in our SGP solution. By iteratively conducting the mask-then-reconstruct operation on incomplete graphs, the latent feature predictor could be regarded as a hidden regularization that assists the graph encoder in extracting more compressed features; 2) RARE consistently outperforms "w/o-Mome." on all eight datasets. Taking the results on Corafull and IMDB-M for example, RARE achieves 3.6% and 2.3% accuracy gains, respectively, demonstrating the effectiveness of providing implicit self-supervision signals to ensure the integrity and accuracy of predicted representations. Similar observations can be concluded from the results on other datasets; and 3) although "w-$\mathcal{F}_m(\mathcal{G})$" can also achieve competitive performance, it suffers from 0.3%–1.3% accuracy degradation compared to our method. The reason behind this may be that since $\mathcal{G}^v$ is a sub-graph of the original graph $\mathcal{G}$, a large amount of redundant information between two-source encoded representations would overwhelm the latent space, resulting in inferior representations for downstream tasks.

TABLE VI
ABLATION STUDY FOR LOSS FUNCTIONS $\mathcal{L}_M$ AND $\mathcal{L}_R$

Method	Loss Function	WikiCS	Flickr	IMDB-M	MUTAG
(A)	$\mathcal{L}_M$ (MSE) & $\mathcal{L}_R$ (MSE)	66.2	49.7	51.0	85.7
(B)	$\mathcal{L}_M$ (ISCE) & $\mathcal{L}_R$ (MSE)	66.5	49.8	50.0	87.3
(C)	$\mathcal{L}_M$ (ISCE) & $\mathcal{L}_R$ (ISCE)	68.2	49.2	51.4	87.2
(D)	$\mathcal{L}_M$ (MAE) & $\mathcal{L}_R$ (ISCE)	68.8	49.9	**53.2**	87.6
Ours	$\mathcal{L}_M$ (MSE) & $\mathcal{L}_R$ (ISCE)	**69.0**	**50.6**	53.1	**88.6**

MAE, MSE, and ISCE are abbreviations for mean absolute error, mean square error, and improved scaled cosine error, respectively. The boldface values indicate the best results (%).

2) Impact of The Loss Function: In this subsection, we conduct ablation studies to investigate the effect of different loss functions. Table VI reports the accuracy results of RARE and its four variants on WikiCS, Flickr, IMDB-M, and MUTAG. $\mathcal{L}_M$ and $\mathcal{L}_R$ indicate the loss functions used for implicit and explicit self-supervision mechanisms, respectively. Moreover, MAE, MSE, and ISCE are abbreviations for mean absolute error, mean square error, and improved scaled cosine error, respectively. From the results presented in Table VI, we can observe that 1) our method achieves better performance than method (A) and method (B) by 2.8%/2.5% and 2.1%/3.1% accuracy improvements on WikiCS and IMDB-M, respectively. The reason behind this is that the MSE loss is better at modeling detailed information from the data itself, while the ISCE loss focuses more on estimating the similarity between two entities. Therefore, minimizing MSE in the noisy raw data space may mislead the network to overly preserve redundant graph details, which may not always result in the required expressive encoding capability for downstream tasks; and 2) RARE and method (D) consistently outperform method (C) by 0.8%/0.6%, 1.4%/0.7%, 1.7%/1.8%, and 1.4%/0.4% on WikiCS, Flickr, IMDB-M, and MUTAG, respectively. This is because the latent features contain much category-related information that needs to be carefully preserved. As a result, completing the masked content in the latent space with MAE or MSE contributes more to the performance than that with ISCE. Moreover, when one has to choose a loss function for masked latent feature completion, both MAE and MSE are suitable for guiding the model learning implicitly.

D. Robustness Against Outliers

To provide a more comprehensive understanding of our motivations, we conduct experiments to make a comparison between

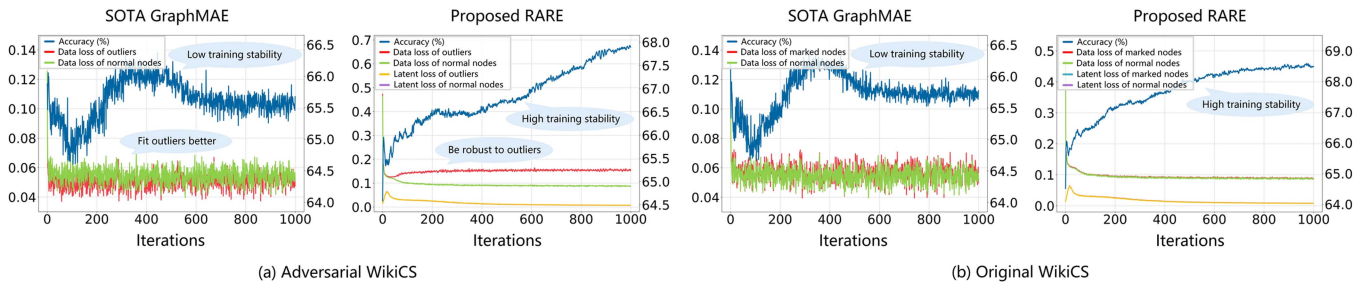

Fig. 4. Method robustness comparison between the proposed RARE and the SOTA GraphMAE [9]: (a) On the adversarial WikiCS; (b) On the original WikiCS. The X-axis, left Y-axis, and right Y-axis refer to the iteration, average loss value, and ACC performance, respectively. Note that the corresponding results on the original WikiCS are provided as a reference. In our setups, both accuracy and loss variations are recorded with iterations. Before pre-training, we randomly generate 5% outliers from the original WikiCS by row-wise shuffling their attributes within the raw attribute matrix. During the pre-training phase, a large random subset of graph nodes (e.g., 75%) is masked out at first and then recovered by minimizing loss values of masked nodes only. After pre-training, we evaluate the model by reporting the performance and the loss curves of normal nodes and all outliers, respectively. As seen, RARE delivers better accuracy and is more robust than GraphMAE [9] in a noisy circumstance.

the proposed RARE and the state-of-the-art (SOTA) GraphMAE on adversarial WikiCS. We also include results from the original WikiCS as a reference. In our setups, we randomly divide all nodes of the original WikiCS into 5% marked nodes and 95% normal nodes. To construct an adversarial WikiCS, within the raw attribute matrix, we row-wise shuffle the attributes of 5% marked nodes to generate outliers. From the sub-figures in Fig. 4, some key observations can be concluded: 1) the proposed RARE achieves an approximate 2.5% accuracy gain against SOTA GraphMAE on both adversarial and original datasets, which demonstrates the superiority and effectiveness of our method; 2) GraphMAE suffers from obvious performance degradation after around 400 iterations until convergence, while RARE substantially enhances the training stability of masked graph autoencoder with performance continually reaching a plateau. We attribute this to the integration of implicit and explicit self-supervision mechanisms, which can regularize each other to provide more reliable guidance for model learning and produce better performance than only minimizing a raw data reconstruction loss; 3) when GraphMAE processes a noisy graph, the average data loss value of outliers (i.e., the red curve) is generally smaller than that of normal nodes (i.e., the green curve), indicating that GraphMAE fits outliers better than normal nodes. This phenomenon is opposite to that of our method, implying that RARE has stronger robustness against adversarial attacks than GraphMAE. These results solidly support our claim that the robustness of the model would be compromised when lacking of reliable self-supervision guidance for model learning; and 4) we also notice an interesting phenomenon that the latent loss curves of outliers and normal nodes almost overlap each other on adversarial WikiCS. We guess that this is because our implicit self-supervision mechanism is driven by the MSE loss, which could ensure a strong alignment of latent features between either outliers or normal samples, leading to subtle distinctions.

E. Hyper-Parameter Sensitivity

1) Impact of Mask Ratio r: To further illustrate the superiority of RARE, we investigate its performance variation with respect to different mask ratios. Concretely, we pre-train RARE by varying the mask ratio r from 0.1 to 0.9 with a step size of 0.1.

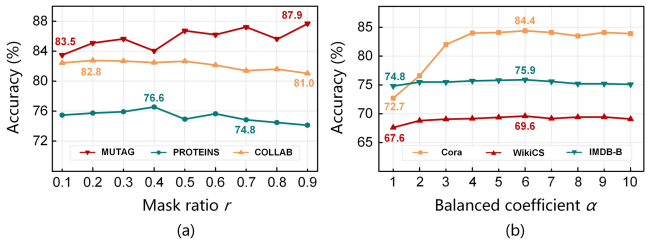

Fig. 5. Performance comparison with the variation of hyper-parameters: (a) The sensitivity of RARE when r varies from 0.1 to 0.9 with 0.1 step size; (b) The sensitivity of RARE when α varies from 1 to 10 with 1 step size. The X-axis and Y-axis refer to the r (or α) value and the ACC performance, respectively.

From the results shown in Fig. 5(a), some key observations can be obtained: 1) taking the results on PROTEINS for example, continually increasing the value of r first improves the model accuracy and then leads to relatively poor performance. This indicates that the mask-then-reconstruct mechanism is indeed effective for self-supervised graph pre-training, but a proper r is required to balance the visible and masked information; 2) increasing r by more than 0.6 would cause a performance drop in most cases, while the proposed RARE can still perform well within a wide range of high mask ratio. For example, the optimal masked ratio (i.e., 90%) for MUTAG is surprisingly high, indicating that in some cases, the model with a high mask ratio largely eliminates redundant information and thus yields a nontrivial and meaningful self-supervision task; and 3) the performance of RARE is relatively stable across a wide range of r on COLLAB. This once again implies that RARE can learn useful representations with limited observed information, indicating its potential to achieve a good accuracy-efficiency trade-off.

2) Impact of Hyper-Parameters α: Eq. (5) introduces a hyper-parameter to balance the importance of two loss functions. To show its influence in-depth, Fig. 5(b) presents the accuracy variation of RARE on three datasets when α varies from 1 to 10 with a step size of 1. Our observations from this sub-figure are as follows: 1) the effect of tuning α on model performance varies across different datasets, but the stability of the model performance is higher in the range of [5, 10]. This indicates that

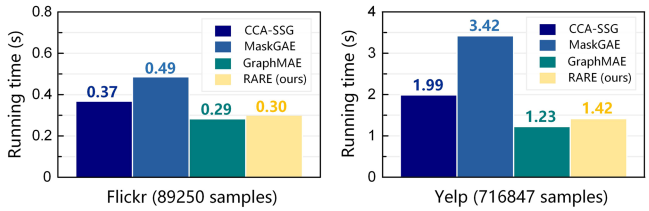

Fig. 6. Running time consumption (second) on Flickr and Yelp. All methods are evaluated on the same device with one NVIDIA 3090 GPU card, and the reported result refers to an average time of 10 iterations.

searching α from a reasonable hyper-parameter region could positively influence the model performance; 2) the accuracy variation is relatively stable across a wide range of α on IMDB-B and WikiCS, while on Cora, it shows a trend of first rising and then dropping slightly. This suggests that RARE requires a suitable α to ensure the quality of learned representations when reconstructing the raw attributes; and 3) RARE tends to perform well by setting $\alpha = 6$ according to the results on three datasets.

F. Running Time Consumption

Fig. 6 shows a comparison of the running time consumption among a scalable contrastive SGP method (i.e., CCA-SSG) and three MGAEs on Flickr and Yelp datasets. All methods are evaluated on the same device with one NVIDIA-3090 GPU card, and the reported result refers to an average time of 10 iterations. As evidenced by the results, since RARE outperforms CCA-SSG in terms of speed on both Flickr and Yelp datasets, the computational efficiency of utilizing a masked graph autoencoder to handle large-scale graphs is promising. Moreover, RARE can achieve better accuracy than MaskGAE and GraphMAE without considerably increasing the computation cost. These experimental results are consistent with previous complexity analyses, once again demonstrating that our method can scale to large-scale scenarios.

More experiments and comparisons are provided in Appendix F and Appendix G, available online, due to the space limit, such as ablation for data-masking setups, ablation for latent feature predictor setups, visualization, and more robustness study.

V. Conclusion and Future Work

In this work, we revisit the inherent distinction between traditional data formats (e.g., images and texts) and graphs for masked signal modeling, and investigate the applicability problem of leveraging masked autoencoders to process graph data. This motivates us to propose a novel framework called RARE for self-supervised graph pre-training. In our method, we implement both implicit and explicit self-supervision mechanisms for masked content recovery by performing a joint mask-then-reconstruct strategy in both latent feature and raw data spaces. Particularly, the designed masked latent feature completion scheme can improve the certainty in inferring masked data and the reliability of the self-supervision mechanism. We also provide theoretical analyses to explain why RARE can work well and how the designed components contribute to the model performance. Extensive experiments on seventeen datasets have demonstrated the effectiveness and superiority of RARE on three downstream tasks.

However, there are still some limitations of existing masked graph autoencoders that have not been fully explored. For instance, existing MGAEs assume that all samples within a graph are available and complete, which may not always hold in practice since it is hard to collect all information from real-world graph data. Future work may extend the proposed RARE to the data-incomplete circumstance, and investigate the connections and differences between self-supervised masked graph pre-training and self-supervised incomplete graph pre-training. Moreover, in the current version, RARE only supports reconstructing each masked node from its adjacent neighbors via graph attention layers. In the future, developing a more effective and efficient MGAE to explore global features for information recovery is another interesting direction.

References

[1] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. B. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15979–15988.
[2] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. 39th Int. Conf. Mach. Learn.*, 2022, pp. 1298–1312.
[3] X. Chen et al., "Context autoencoder for self-supervised representation learning," *Int. J. Comput. Vis.*, 2023.
[4] Z. Chen et al., "Multi-modal masked autoencoders for medical vision-and-language pre-training," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2022, pp. 679–689.
[5] Z. Fu, C. Wang, J. Xu, H. Zhou, and L. Li, "Contextual representation learning beyond masked language modeling," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 2701–2714.
[6] A. Chen et al., "PiMAE: Point cloud and image interactive masked autoencoders for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5291–5301.
[7] Q. Tan et al., "S2GAE: Self-supervised graph autoencoders are generalizable learners with graph masking," in *Proc. 16th ACM Int. Conf. Web Search Data Mining*, 2023, pp. 787–795.
[8] J. Li et al., "What's behind the mask: Understanding masked graph modeling for graph autoencoders," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2023, pp. 1268–1279.
[9] Z. Hou et al., "GraphMAE: Self-supervised masked graph autoencoders," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2022, pp. 594–604.
[10] Y. Tian, K. Dong, C. Zhang, C. Zhang, and N. V. Chawla, "Heterogeneous graph masked autoencoders," in *Proc. 37th AAAI Conf. Artif. Intell.*, 2023, pp. 9997–10005.
[11] C. Liu, Y. Zhan, X. Ma, D. Tao, B. Du, and W. Hu, "Masked graph auto-encoder constrained graph pooling," in *Proc. Mach. Learn. Knowl. Discov. Databases: Eur. Conf.*, 2022, pp. 377–393.
[12] Z. Wang et al., "BatmanNet: Bi-branch masked graph transformer autoencoder for molecular representation," 2022, *arXiv:2211.13979*.
[13] X. Li, T. Ye, C. Shan, D. Li, and M. Gao, "SeeGera: Self-supervised semi-implicit graph variational auto-encoders with masking," in *Proc. ACM Web Conf.*, 2023, pp. 143–153.
[14] W. Yu, M. Huang, S. Wu, and Y. Zhang, "Ensembled masked graph autoencoders for link anomaly detection in a road network considering spatiotemporal features," *Inf. Sci.*, vol. 622, pp. 456–475, 2023.
[15] J. Feng, Z. Wang, Y. Li, B. Ding, Z. Wei, and H. Xu, "MGMAE: Molecular representation learning by reconstructing heterogeneous graphs with a high mask ratio," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manage.*, 2022, pp. 509–519.
[16] K. Jing, J. Xu, and P. Li, "Graph masked autoencoder enhanced predictor for neural architecture search," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, 2022, pp. 3114–3120.
[17] C. Morris et al., "Weisfeiler and leman go neural: Higher-order graph neural networks," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, pp. 4602–4609.

[18] Z. Lin, Z. Kang, L. Zhang, and L. Tian, "Multi-view attributed graph clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 2, pp. 1872–1880, Feb. 2023.
[19] J. Li et al., "Higher-order attribute-enhancing heterogeneous graph neural networks," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 560–574, Jan. 2023.
[20] X. Zhao, B. Zong, Z. Guan, K. Zhang, and W. Zhao, "Substructure assembling network for graph classification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4514–4521.
[21] W. Lu et al., "SkipNode: On alleviating performance degradation for deep graph convolutional networks," 2021, *arXiv:2112.11628*.
[22] Y. Liu et al., "Graph self-supervised learning: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 5879–5900, Jun. 2023.
[23] Y. Yang, Z. Guan, J. Li, W. Zhao, J. Cui, and Q. Wang, "Interpretable and efficient heterogeneous graph convolutional network," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 2, pp. 1637–1650, Feb. 2023.
[24] Y. Yang, Z. Guan, W. Zhao, W. Lu, and B. Zong, "Graph substructure assembling network with soft sequence and context attention," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 5, pp. 4894–4907, May 2023.
[25] J. Zhao et al., "IntentGC: A scalable graph convolution framework fusing heterogeneous information for recommendation," in *Proc. 25th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2019, pp. 2347–2357.
[26] M. You, A. Yuan, M. Zou, D. He, and X. Li, "Robust unsupervised feature selection via multi-group adaptive graph representation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 3030–3044, Mar. 2023.
[27] L. Gong, W. Tu, S. Zhou, L. Zhao, Z. Liu, and X. Liu, "Deep fusion clustering network with reliable structure preservation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Nov. 17, 2022, doi: 10.1109/TNNLS.2022.3220914.
[28] S. Wang et al., "Highly-efficient incomplete large-scale multi-view clustering with consensus bipartite graph," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9776–9785.
[29] K. Liang et al., "Learn from relational correlations and periodic events for temporal knowledge graph reasoning," in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2023, pp. 1559–1568.
[30] K. Liang et al., "Knowledge graph contrastive learning based on relation-symmetrical structure," *IEEE Trans. Knowl. Data Eng.*, early access, Jun. 12, 2023, doi: 10.1109/TKDE.2023.3282989.
[31] P. Velickovic, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," in *Proc. Int. Conf. Learn. Representations*, 2019.
[32] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 5812–5823.
[33] J. Qiu et al., "GCC: Graph contrastive coding for graph neural network pre-training," in *Proc. 26th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2020, pp. 1150–1160.
[34] K. Hassani and A. H. Khasahmadi, "Contrastive multi-view representation learning on graphs," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 4116–4126.
[35] L. Gong, S. Zhou, W. Tu, and X. Liu, "Attributed graph clustering with dual redundancy reduction," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 3015–3021.
[36] F.-Y. Sun, J. Hoffman, V. Verma, and J. Tang, "InfoGraph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization," in *Proc. Int. Conf. Learn. Representations*, 2020.
[37] Y. Liu et al., "Hard sample aware network for contrastive deep graph clustering," in *Proc. 37th AAAI Conf. Artif. Intell.*, 2023, pp. 8914–8922.
[38] Z. Hou, Y. Cen, Y. Dong, J. Zhang, and J. Tang, "Automated unsupervised graph representation learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 2285–2298, Mar. 2023.
[39] T. N. Kipf and M. Welling, "Variational graph auto-encoders," 2016, *arXiv:1611.07308*.
[40] W. Tu et al., "Deep fusion clustering network," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 9978–9987.
[41] Y. Lu, X. Jiang, Y. Fang, and C. Shi, "Learning to pre-train graph neural networks," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 4276–4284.
[42] Y. Liu et al., "Deep graph clustering via dual correlation reduction," in *Proc. 36th AAAI Conf. Artif. Intell.*, 2022, pp. 7603–7611.
[43] Y. Yang et al., "Self-supervised heterogeneous graph pre-training based on structural clustering," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 16962–16974.
[44] X. Gao, W. Hu, and G.-J. Qi, "Self-supervised graph representation learning via topology transformations," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 4202–4215, Apr. 2023.
[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.
[46] S. Zhang, H. Chen, H. Yang, X. Sun, P. S. Yu, and G. Xu, "Graph masked autoencoders with transformers," 2022, *arXiv:2202.08391*.
[47] S. Yun, M. Jeong, R. Kim, J. Kang, and H. J. Kim, "Graph transformer networks," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 11960–11970.
[48] L. Xia, C. Huang, C. Huang, K. Lin, T. Yu, and B. Kao, "Automated self-supervised learning for recommendation," in *Proc. ACM Web Conf.*, 2023, pp. 992–1002.
[49] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.
[50] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 139–156.
[51] J.-B. Grill et al., "Bootstrap your own latent - A new approach to self-supervised learning," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.
[52] S. Thakoor et al., "Large-scale representation learning on graphs via bootstrapping," in *Proc. Int. Conf. Learn. Representations*, 2021.
[53] Y. Liu, Y. Zheng, D. Zhang, H. Chen, H. Peng, and S. Pan, "Towards unsupervised deep graph structure learning," in *Proc. ACM Web Conf.*, 2022, pp. 1392–1403.
[54] N. Lee, J. Lee, and C. Park, "Augmentation-free self-supervised learning on graphs," in *Proc. 36th AAAI Conf. Artif. Intell.*, 2022, pp. 7372–7380.
[55] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018.
[56] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," in *Proc. Int. Conf. Learn. Representations*, 2019.
[57] Q. Zhang, Y. Wang, and Y. Wang, "How mask matters: Towards theoretical understandings of masked autoencoders," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 27127–27139.
[58] J. Zhu, L. Tao, H. Yang, and F. Nie, "Unsupervised optimized bipartite graph embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 3224–3238, Mar. 2023.
[59] L. Wu, H. Lin, C. Tan, Z. Gao, and S. Z. Li, "Self-supervised learning on graphs: Contrastive, generative, or predictive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 4216–4235, Apr. 2023.
[60] Y. LeCun et al., "Handwritten zip code recognition with multilayer networks," in *Proc. 10th Int. Conf. Pattern Recognit.*, 1990, pp. 35–40.
[61] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffne, "Gradient-based learning applied to document recognition," in *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[62] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.
[63] H. Zeng, H. Zhou, A. Srivastava, R. Kannan, and V. K. Prasanna, "Graph-SAINT: Graph sampling based inductive learning method," in *Proc. Int. Conf. Learn. Representations*, 2020.
[64] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017.
[65] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, and L. Wang, "Deep graph contrastive representation learning," 2020, *arXiv:2006.04131*.
[66] D. Xu, W. Cheng, D. Luo, H. Chen, and X. Zhang, "InfoGCL: Information-aware graph contrastive learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 30414–30425.
[67] H. Zhang, Q. Wu, J. Yan, D. Wipf, and P. S. Yu, "From canonical correlation analysis to self-supervised graph neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 76–89.
[68] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, "Weisfeiler-lehman graph kernels," *J. Mach. Learn. Res.*, vol. 12, no. 9, pp. 2539–2561, 2011.
[69] P. Yanardag and S. Vishwanathan, "Deep graph kernels," in *Proc. 21th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2015, pp. 1365–1374.
[70] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable poolings," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 4805–4815.
[71] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal, "graph2vec: Learning distributed representations of graphs," 2017, *arXiv:1707.05005*.
[72] Y. You, T. Chen, Y. Shen, and Z. Wang, "Graph contrastive learning automated," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 12121–12132.

[73] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.
[74] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

Wenxuan Tu is currently working toward the PhD degree with the School of Computer, National University of Defense Technology (NUDT), Changsha, China. His research interests include clustering analysis, graph machine learning, and image semantic segmentation. He has published several papers in highly regarded journals and conferences, such as *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Neural Networks and Learning Systems*, AAAI, IJCAI, CVPR, NeurIPS, ICML, ACM MM, SIGIR, etc.

Qing Liao received the PhD degree in computer science and engineering from the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, in 2016 supervised by Prof. Qian Zhang. She is currently a professor with the College of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), China. Her research interests include artificial intelligence and data mining.

Sihang Zhou received the bachelor's degree in information and computing science from the University of Electronic Science and Technology of China (UESTC), in 2012, and the MS degree in computer science from the National University of Defense Technology (NUDT), China, in 2014, and the PhD degree from the National University of Defense Technology (NUDT), China, in 2019. He is now an associate professor with the College of Intelligence Science and Technology, NUDT. His current research interests include machine learning, knowledge graph, and medical image analysis.

Xin Peng received the MS degree from Hainan University, in 2023. He is currently working toward the PhD degree with the School of Computer, National University of Defense Technology (NUDT), Changsha, China. His research interests include graph representation learning and semantic segmentation. He has published several papers in highly regarded journals, such as *IEEE Transactions on Neural Networks and Learning Systems*, Wiley *International Journal of Information Security*, and *Information Fusion*.

Chuan Ma (Member, IEEE) received the BS degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2013 and PhD degree from the University of Sydney, Australia, in 2018. From 2018 to 2022, he worked as a lecturer with the Nanjing University of Science and Technology, and now he is a principal investigator with Zhejiang Lab. He has published more than 40 journal and conference papers, including a best paper in WCNC 2018, and a best paper award in IEEE Signal Processing Society 2022. His research interests include stochastic geometry, wireless caching networks, and distributed machine learning, and he now focuses on Big Data analysis and privacy-preserving.

Zhe Liu (Senior Member, IEEE) received the BS and MS degrees from Shandong University, in 2008 and 2011, respectively, and the PhD degree from the Laboratory of Algorithmics, Cryptology, and Security (LACS), University of Luxembourg, Luxembourg, in 2015. He is a professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA), China, and he is the founder of the Trust Intelligent Computing Research Group, Zhejiang Lab. Before joining NUAA, he was a researcher with the SnT, University of Luxembourg, Luxembourg. His PhD thesis has received the prestigious FNR Awards 2016—Outstanding PhD Thesis Award for his contributions to cryptographic engineering on IoT devices. His research interests include computer arithmetic and information security. He has co-authored more than 70 research peer-reviewed journal and conference papers.

Xinwang Liu (Senior Member, IEEE) received the PhD degree from the National University of Defense Technology (NUDT), China. He is now a full professor with the College of Computer, NUDT. His current research interests include kernel learning and unsupervised feature learning. He has published more than 100 peer-reviewed papers, including those in highly regarded journals and conferences such as *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Information Forensics and Security*, ICML, NeurIPS, ICCV, CVPR, AAAI, IJCAI, etc. He serves as the associated editor of *Information Fusion Journal*, *IEEE Transactions on Cybernetics*, and *IEEE Transactions on Neural Networks and Learning Systems* Journal.

Zhiping Cai received the BS, MS, and PhD degrees in computer science and technology from the National University of Defense Technology (NUDT), Changsha, China, in 1996, 2002, and 2005, respectively. He is a full professor with the College of Computer, NUDT. His current research interests include artificial intelligence, network security, and Big Data. He is a senior member of CCF.

Kunlun He received the MD degree from The 3rd Military Medical University, Chongqing, China, in 1988, and the PhD degree in cardiology from Chinese PLA Medical School, Beijing, China, in 1999. He worked as a postdoctoral research fellow with the Division of Circulatory Physiology, Columbia University from 1999 to 2003. He is the director and professor of the Medical Big Data Research Center, Chinese PLA General Hospital, Beijing, China. His research interests include Big Data and artificial intelligence of cardiovascular disease.