

Cross-Domain Graph Anomaly Detection via Test-Time Training with Homophily-Guided Self-Supervision

Anonymous authors

Paper under double-blind review

Abstract

Graph Anomaly Detection (GAD) has demonstrated great effectiveness in identifying unusual patterns within graph-structured data. However, while labeled anomalies are often scarce in emerging applications, existing supervised GAD approaches are either ineffective or not applicable when moved across graph domains due to distribution shifts and heterogeneous feature spaces. To address these challenges, we present GADT3, a novel test-time training framework for cross-domain GAD. GADT3 combines supervised and self-supervised learning during training while adapting to a new domain during test time using only self-supervised learning by leveraging a homophily-based affinity score that captures domain-invariant properties of anomalies. Our framework introduces four key innovations to cross-domain GAD: an effective self-supervision scheme, an attention-based mechanism that dynamically learns edge importance weights during message passing, domain-specific encoders for handling heterogeneous features, and class-aware regularization to address imbalance. Experiments across multiple cross-domain settings demonstrate that GADT3 significantly outperforms existing approaches, achieving average improvements of over 8.2% in AUROC and AUPRC compared to the best competing model.

1 Introduction

Graph Anomaly Detection (GAD) is a critical task to identify unusual patterns (or outliers) in graph-structured data (Ma et al., 2021; Akoglu et al., 2015). This problem has many real-world applications such as in e-commerce (Zhang et al., 2022), social networks (Venkatesan & Prabhavathy, 2019), fraud detection (Jiang et al., 2019), and cybersecurity (Lazarevic et al., 2003). A key limitation of existing GAD models is that they face unique cross-domain challenges that distinguish them from general graph learning tasks. These include inconsistent definitions of normal/anomalous patterns and heterogeneous feature spaces between domains (e.g., Facebook vs. Amazon datasets). These applications would benefit from GAD models capable of adapting across Out-Of-Distribution (OOD) and heterogeneous graphs.

A major motivation for our work is cybersecurity. Network intrusion detection systems play a key role in identifying malicious network activity associated with cyberattacks (Tsai et al., 2009; Kilincer et al., 2021). Due to the volume and dynamic nature of these attacks, modern intrusion detection systems increasingly rely on large amounts of data and machine learning to assist cybersecurity experts in detecting potential intrusions. However, one of the major challenges in data-driven cybersecurity is the lack of sufficient labeled data for training supervised models. This is particularly critical for emerging applications, where traces of labeled intrusions are scarce and the ability to leverage attacks or other types of anomalous behavior from existing labeled datasets would be greatly beneficial.

This paper investigates the unsupervised domain adaptation problem for GAD focusing on the transfer of knowledge from a labeled source to an unlabeled target domain. While domain adaptation for graphs has been studied by previous work (Shi et al., 2024; Wang et al., 2021; Zheng et al., 2019; Yin et al., 2023; Wang et al., 2023a; Ding et al., 2018), cross-domain graph anomaly detection is an emerging challenge (Wang et al., 2023b; Ding et al., 2021). We propose using Test-Time Training (TTT) (Sun et al., 2020) to address cross-domain graph anomaly detection. Unlike traditional domain adaptation methods that require continuous

access to source data, TTT allows us to distill source knowledge into model parameters and adapt to target domains using only self-supervision. This is particularly advantageous for graph anomaly detection where source data may be sensitive or unavailable during the adaptation phase.

Our approach, Graph Anomaly Detection with Test-time Training (GADT3), has four major innovations compared to existing work on cross-domain GAD. First, it leverages a homophily-based affinity score (Qiao & Pang, 2024; Chen et al., 2024) for self-supervised learning based on the observation that normal nodes are more similar to their neighbors than anomalous ones, a pattern that remains consistent across domains. We observe a clear separation in homophily scores between normal and anomalous nodes across most datasets, justifying the use of homophily-based SSL. Figure 1 shows the homophily score distributions for two datasets (see Appendix A.8 for detailed plots). Second, it introduces Normal Structure-preserved Attention Weighting (NSAW), which dynamically learns continuous edge importance weights through the attention mechanism to suppress the effect of anomalous nodes on the representation of normal ones. Third, GADT3 applies source and target-specific encoders that are trained end-to-end to handle both distribution shifts and heterogeneous feature spaces. Fourth, to address the extreme class imbalance challenge in GAD (Ma et al., 2024), our approach employs class-aware regularization during source training—with a stronger regularization to the minority class. By combining these innovations, GADT3 is able to effectively identify anomalous nodes across heterogeneous datasets arising from multiple and diverse domains.

We compare our solution against both graph domain adaptation and graph anomaly detection approaches using multiple cross-domain datasets. For instance, we show that labeled anomalies in the Amazon dataset (source) can improve the GDA accuracy on the Facebook dataset (target). The experiments show that GADT3 significantly outperforms the alternatives in most of the settings. The contributions of this paper can be summarized as:

- We propose the first test-time training (TTT) framework for cross-domain graph anomaly detection (GADT3). TTT enables adaptation to new target domains without requiring direct access to source data or target labels at test time. Our approach leverages homophily-based self-supervision which captures universal patterns, as anomalous nodes typically exhibit lower homophily compared to normal nodes.
- Our message-passing technique introduces homophily-based attention weights that naturally suppress anomalous influences on normal nodes while preserving graph structure, enabling robust anomaly detection.
- We propose several strategies specifically designed for cross-domain GAD such as data-specific encoders, class-aware regularization, and an early-stopping strategy to prevent overfitting during test-time adaptation.
- GADT3’s performance is demonstrated through extensive experiments on cross-network tasks using multiple datasets and baselines that span both domain adaptation and anomaly detection. Our approach achieves average relative improvements of over 8.2% in AUROC and AUPRC compared to the best competing model.

2 Related Work

2.1 Graph Domain Adaptation (GDA)

While graph domain adaptation (GDA) methods have shown success in transferring knowledge across different domains (Shi et al., 2024), they face fundamental limitations when applied to anomaly detection tasks. As traditional GDA methods are designed for node classification scenarios with balanced classes and consistent label semantics.

Traditional Graph Domain Adaptation (GDA) approaches, such as GRADE (Wu et al., 2023), AdaGCN (Dai et al., 2022), and UDA-GCN (Wu et al., 2020) and spectral regularization methods (You et al., 2023) focus on minimizing distribution shifts across domains but require training samples during inference, limiting their

real-time adaptation capabilities. Test-time training (TTT) (Sun et al., 2020) has emerged as a powerful framework for handling distribution shifts (Li et al., 2022; Wu et al., 2024; Zhang et al., 2024; Liu et al., 2021b). Methods like GTrans (Jin et al., 2022), GraphCL (You et al., 2020), and GT3 (Wang et al., 2022) propose graph-specific TTT approaches. TENT (Wang et al., 2020) uses prediction entropy minimization, while GraphTTA (Chen et al., 2022) leverages information theory for TTT on graphs. However, existing GDA approaches, including both traditional methods and TTT-based solutions, face key limitations for anomaly detection as they (1) don’t account for extreme class imbalance within domains, (2) lack mechanisms to preserve anomaly-indicating patterns during transfer, and (3) don’t handle heterogeneous feature spaces across domains. While there exists limited research on cross-domain graph anomaly detection, prior approaches require direct access to the source dataset during adaptation (Ding et al., 2021; Wang et al., 2023b). Our approach offers key advantages over these methods by eliminating the need to store and access the full source dataset during adaptation. This makes GADT3 both privacy-preserving and memory-efficient, making it more practical for real-world deployment where data privacy and computational resources are concerns.

Our work addresses key challenges in deploying GDA in anomaly detection through dataset-specific encoders for heterogeneous feature spaces and a homophily-based unsupervised learning approach that extracts domain-invariant properties of anomalies that generalize across different graph domains. In our experiments, we show that GADT3 often outperforms multiple GDA baselines, including GRADE, AdaGCN, TENT, and GTrans.

2.2 Graph Anomaly Detection (GAD)

Graph Anomaly Detection (GAD) (Ma et al., 2021; Akoglu et al., 2015) focuses on identifying abnormal nodes in graph-structured data. Traditional approaches like Oddball (Akoglu et al., 2010) rely on power-law relationships between local graph features, while more recent deep learning-based approaches are more generalizable. For instance, DOMINANT (Ding et al., 2019) employs a graph autoencoder to identify anomalies based on graph reconstruction. ComGA (Luo et al., 2022) introduces a tailored GCN to learn distinguishable node representations by explicitly capturing community structure. Self-supervised techniques have emerged as powerful tools for GAD, with methods like CoLA (Liu et al., 2021a), SL-GAD (Zheng et al., 2021), HCM-A (Huang et al., 2022), and TAM (Qiao & Pang, 2024) introducing various approaches to handle node interactions and structural patterns. While these methods have shown success in single-domain scenarios, they don’t address the challenges of cross-domain knowledge transfer. Recent methods like ACT (Wang et al., 2023b) and COMMANDER (Ding et al., 2021) have started to explore cross-domain graph anomaly detection (CD-GAD), but this remains an evolving field. Anomalies behave differently across domains—fraudulent users in e-commerce networks exhibit different patterns compared to social networks. Second, domains often have different feature spaces and graph structures.

GADT3 bridges both GDA and GAD through a unified approach that combines (1) test-time training with a homophily-based loss for capturing domain-invariant properties of anomalies (2) enhanced message-passing using normal structure-preserved attention weighting (NSAW) that reduces the irrelevant impact of anomalous nodes on normal nodes’ representations, and (3) class-aware regularization that prevents minority class patterns from being overshadowed by majority class during source training. We compare a source-free version of GADT3 against GAD baselines (e.g. TAM, ComGA, DOMINANT) to highlight the impact of its GAD-specific features.

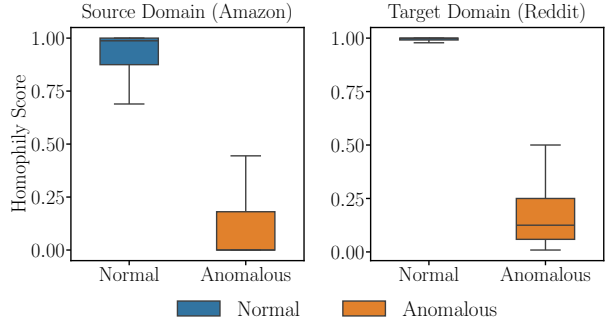


Figure 1: Homophily score distributions across domains (Amazon to Reddit): normal nodes (blue) show higher scores than anomalous nodes (orange), suggesting homophily as a domain-invariant anomaly signal.

3 Problem Definition

We address unsupervised node-level anomaly detection with domain adaptation, aiming to identify abnormal nodes in a target graph by leveraging information from both the target and a source graph. The source dataset contains labeled information but is assumed to be out-of-distribution (OOD) relative to the target, with different feature sets and label distributions. Given source dataset $D_s = (\mathcal{G}_s, X_s, Y_s)$ and target dataset $D_t = (\mathcal{G}_t, X_t)$, where $\mathcal{G}_s/\mathcal{G}_t = (\mathcal{V}, \mathcal{E})$ represents a (source or target) underlying graph with nodes \mathcal{V} and edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ such that $uv \in \mathcal{E}$ if there is a link between nodes u and v . Here, $X_s = \{\mathbf{x}_v | \forall v \in \mathcal{V}_s\}$ and $Y_s = \{y_v | \forall v \in \mathcal{V}_s\}$ represent the node features and their corresponding labels, respectively, and the same holds for target data (X_t, Y_t) . Notably, $y_v \in \{0, 1\}$ (normal or anomaly) for nodes in both \mathcal{V}_s and \mathcal{V}_t . Our goal is to detect anomalies in the target graph. The key challenge is that target labels are completely unavailable (unsupervised setting), requiring us to leverage source labels Y_s as auxiliary information, even though source and target features may have different dimensionalities and semantic meanings ($\mathbf{x}_v \in \mathbb{R}^{P_s}$ for source nodes and $\mathbf{x}_u \in \mathbb{R}^{P_t}$ for target nodes).

4 Graph Anomaly Detection via Test-Time Training (GADT3)

We propose GADT3 (Graph Anomaly Detection with Test-time Training), a domain adaptation method particularly designed for graph anomaly detection that integrates attention-based message-passing, node homophily patterns, and domain-specific encoders to handle feature shifts and heterogeneous features across domains. GADT3 addresses the challenging and novel scenario where the target data is out of distribution (OOD) relative to the source data and where source and target datasets have different feature spaces.

An overview of our approach is provided in Figure 2. GADT3 leverages the advantages of Test-time Training, especially the ability to adapt to new target datasets in the TTT phase without the need for labeled anomalies. Moreover, GADT3 incorporates technical innovations tailored for cross-domain GAD, including the use of domain-specific encoders, Normal Structure-preserved Attention Weighting (NSAW), and class-aware regularization. Unlike prior TTT methods that rely on auxiliary dummy tasks (e.g., rotation prediction or jigsaw solving) for encoder adaptation, our self-supervised homophily loss is directly aligned with the target anomaly detection objective. By encouraging high homophily among node embeddings in the target graph, this task-relevant signal promotes local consistency, enabling the model to adapt without labels while preserving the semantic structure essential for anomaly detection.

Our framework operates in two phases. The *training phase* combines supervised and self-supervised losses to learn from the source dataset. The *test-time training phase* only uses self-supervised learning to adapt to the target data. During training, we learn the encoder θ_s , decoder θ_m , and supervised predictor θ_{pred} jointly based on the source dataset. During test-time training (TTT) on the target domain, a new encoder θ_t is trained from scratch using the unsupervised homophily-based loss $\mathcal{L}_{\text{self}}$, while θ_m is reused from the training phase. This decoupled encoder-decoder structure enables adaptation without requiring any target labels or access to source data after training. We will detail each stage of our solution together with our key contributions in the next sections.

4.1 Backbone GNN and Projection head

The backbone architecture of GADT3 is a Graph Neural Network (GNN) that operates on source and target graph data. Our model can apply any GNN architecture, including Message-passing Neural Networks (MPNNs). We denote the set of neighbors at node v as $\mathcal{N}(v) = \{u | (u, v) \in \mathcal{E}\}$. For each layer $\ell \in \{1, 2, \dots, L\}$ of the GNN the information from neighboring nodes of v is aggregated via message-passing as follows:

$$\mathbf{h}_{\mathcal{N}(v)}^\ell = M(\{\mathbf{h}_u^{\ell-1} | u \in \mathcal{N}(v)\}), \quad (1)$$

where $M(\cdot)$ is an element-wise and permutation-invariant operator (e.g., average, max, or min). The embedding of node v at the ℓ -th layer is derived from the aggregated neighbors' embeddings and the previous embedding of v :

$$\mathbf{h}_v^\ell = \sigma\left(W^\ell \cdot [\mathbf{h}_{\mathcal{N}(v)}^\ell || \mathbf{h}_v^{\ell-1}] + \mathbf{b}^\ell\right), \quad (2)$$

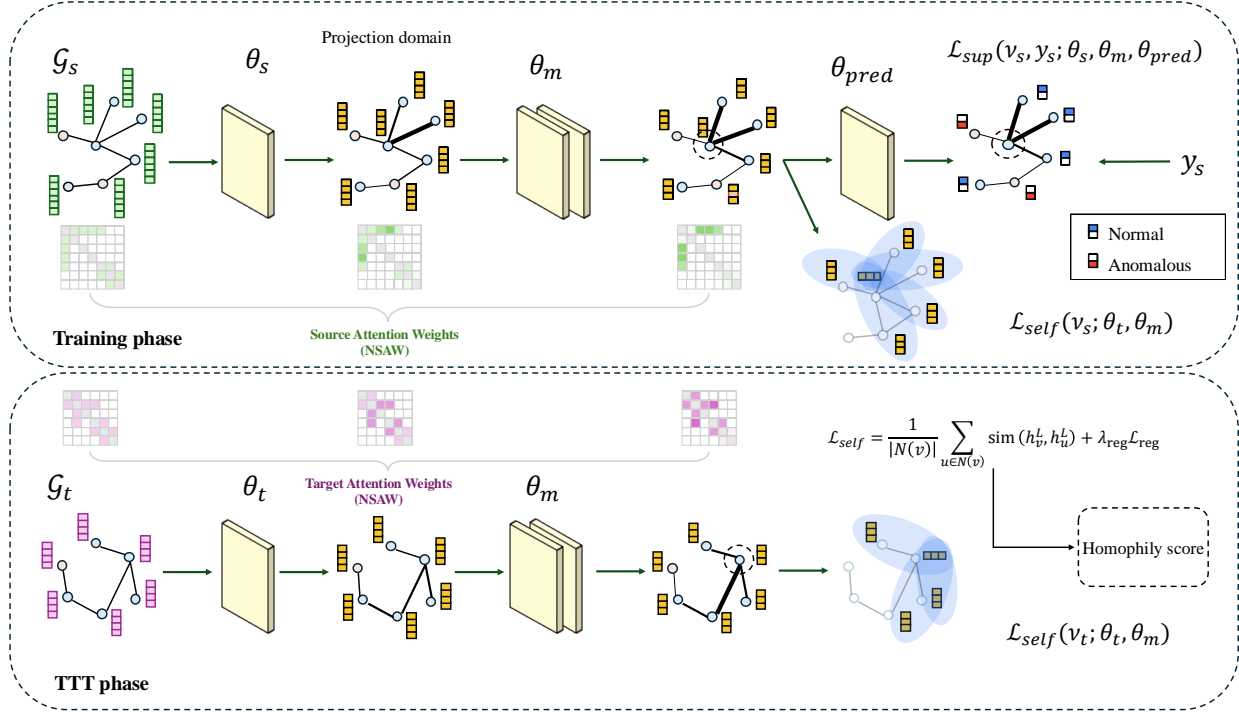


Figure 2: Overview of GADT3, our test-time training framework for node-level cross-domain anomaly detection. In the training phase (top), GADT3 learns an encoder θ_s , a decoder θ_m , and a prediction θ_{pred} by jointly minimizing supervised (\mathcal{L}_{sup}) and self-supervised homophily-based loss (\mathcal{L}_{self}). We propose using Normal Structure-preserved Attention Weights to defend against the adversarial influence of anomalous nodes during message passing in the GNN training. In the test-time training (TTT) phase (bottom), GADT3 learns a target encoder θ_t using only the self-supervised loss \mathcal{L}_{self} . The decoder θ_m is shared across phases to enable the cross-domain knowledge transfer while handling heterogeneous feature spaces. During inference, the adapted target representations produced by θ_m are used for anomaly scoring.

where $\sigma(\cdot)$ is a non-linear activation function, \parallel denotes the concatenation operator, $W^\ell \in \mathbb{R}^{p_{\ell+1} \times 2p_\ell}$, and $\mathbf{b}^\ell \in \mathbb{R}^{p_{\ell+1}}$ are trainable weight matrices and bias vectors. Representations from the last layer ($\ell = L$) can be converted to binary class probabilities using the sigmoid function. To generate initial representations \mathbf{h}_v^0 for nodes with different feature sets in D_s and D_t , GADT3 learns projection matrices (or encoders) P_s and P_t as follows:

$$\mathbf{h}_v^0 = P_s \cdot \mathbf{x}_v, \quad (3)$$

where $P_s \in \mathbb{R}^{p \times p_s}$ is a trainable projection for the source data and \mathbf{x}_v are node features. The same is performed for target data using the matrix P_t . These encoders enable projecting source and target data into a shared representation space, facilitating transfer learning.

Without loss of generality, we apply GraphSAGE (Hamilton et al., 2017) as GADT3’s backbone MPNN architecture. GraphSAGE applies node sampling in the aggregation phase to increase its scalability.

4.2 Homophily-based Self-supervised Loss

Test-time training requires a self-supervised loss \mathcal{L}_{self} applied during both training and test-time training stages. We propose the use of a previously introduced homophily-based affinity score, which is tailored for unsupervised anomaly detection (Qiao & Pang, 2024; Chen et al., 2024).

The local affinity score measures the similarity or connection strength between a node and its neighbors. It is based on the one-class homophily phenomenon, where normal nodes tend to have a stronger affinity with their neighbors compared with anomalies. This enables these scores to be applied for unsupervised GAD. To the best of our knowledge, our work is the first to apply affinity scores for the GAD problem within the test-time training framework.

More formally, the local affinity score is computed by comparing a node’s representation against those of its immediate neighbors in the graph using various similarity metrics such as cosine similarity, Euclidean distance, Jensen-Shannon divergence, and Wasserstein distance (Lin, 1991; Chen & Liu, 2020; Shen et al., 2018). By taking the average similarity with neighbors of a node v , we obtain a single anomaly score $s(v)$ based on how well a node is associated with its local graph structure:

$$s(v) = \frac{1}{|N(v)|} \sum_{u \in N(v)} \text{sim}(\mathbf{h}_v^L, \mathbf{h}_u^L), \quad (4)$$

where $N(v_i)$ denotes the neighbors of node v_i , \mathbf{h}_u^L represents the learned node embedding for u , and $\text{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^T \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$ describes the cosine similarity. The unsupervised task consists of maximizing the affinity score for each node.

This test-time task is scalable, as it only requires local computations, making it efficient even for large graphs. Additionally, it can be adapted to different types of graphs and node features by choosing appropriate node representation learning techniques and similarity measures. The total self-supervised loss combines the homophily-based loss with a regularization term:

$$\mathcal{L}_{\text{self}} = - \sum_{v \in \mathcal{V}} s(v) + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}, \quad (5)$$

where $\mathcal{L}_{\text{reg}} = \sum_{v \in \mathcal{V}} \left(\frac{\sum_{u \in \mathcal{N}_v^-} \text{sim}(\mathbf{h}_v, \mathbf{h}_u)}{|\mathcal{N}_v^-|} \right)$ minimizes the similarity between non-connected nodes to maintain structural distinctiveness in the embedding space, with \mathcal{N}_v^- representing the set of nodes not connected to node v and λ controlling the regularization strength.

This homophily loss also serves as an implicit alignment mechanism, encouraging local consistency among neighboring node embeddings in the target graph. By leveraging the frozen source decoder θ_m , the model preserves the source-learned graph patterns, achieving alignment without requiring explicit alignment terms or target labels.

It has been noted by previous work (Qiao & Pang, 2024; Chen et al., 2024) that enforcing similarity across all connected nodes might inadvertently cause non-homophily nodes to become similar. Normal Structure-preserved Graph Truncation (NSGT) attempts to preserve normal node structure by making binary decisions to remove edges between dissimilar nodes based on distance. However, this approach risks losing important structural information and requires setting explicit thresholds for edge removal decisions. In the next section, we introduce our enhanced message-passing approach through attention-based NSAW, a simpler and more flexible approach to suppress the adversarial effect of non-homophily edges in equation 5.

4.3 Normal Structure-preserved Attention Weighting (NSAW)

We propose Normal Structure-preserved Attention Weighting (NSAW) as an alternative to Normal Structure-preserved Graph Truncation (Qiao & Pang, 2024). NSAW applies a robust extension to graph attention (Velićković et al., 2017) as a more flexible way to suppress the effect of anomalous nodes on the representation of normal nodes during message-passing. These attention weights are learned end-to-end as part of the GNN training process.

For each layer ℓ , we compute attention weights between connected nodes using a learnable transformation matrix $U^\ell \in \mathbb{R}^{p_\ell \times \tilde{p}}$, where \tilde{p} represents the attention dimension. Given representations $H^\ell \in \mathbb{R}^{N \times p_\ell}$ at layer

ℓ , the attention scores are computed as:

$$A^\ell = \text{softmax}(\phi(H^\ell U^\ell)(\phi(H^\ell U^\ell))^\top \odot M), \quad (6)$$

where ϕ is ReLU activation, M is the adjacency matrix, and \odot is element-wise multiplication. The use of M ensures that attention weights are only computed between connected nodes in the graph.

Robust Symmetric Attention. To defend against the adversarial influence of anomalous nodes, we enforce symmetry in the attention weights through a symmetric minimum operation:

$$\tilde{A}_{i,j}^\ell = \min\{A_{i,j}^\ell, A_{j,i}^\ell\}, \quad \text{for all } i \neq j \quad (7)$$

NSAW’s symmetric construction provides a crucial defensive mechanism through two complementary aspects. First, when a normal node v connects to both normal nodes $\{v_1, v_2, \dots, v_k\}$ and anomalous nodes $\{a_1, a_2, \dots, a_m\}$, the attention naturally assigns high weights to normal neighbors and low weights to anomalous ones due to feature similarity. As illustrated in the Figure 3, when a normal node assigns low attention $\alpha(v \rightarrow a) \approx 0$ to an anomalous node (thin arrow), even if the anomalous node attempts to assign high attention back (thick arrow), the minimum operation then ensures that the reverse influence is also minimized as $\alpha(a \rightarrow v) = \min(\alpha(a \rightarrow v), \alpha(v \rightarrow a)) \approx 0$ (shown by the dashed arrow), regardless of the original attention weight. Second, through the learnable attention parameters U^l and the self-supervised loss, the architecture gradually learns to suppress the influence of anomalous patterns during training, further strengthening this defensive mechanism through the GNN’s message passing structure.

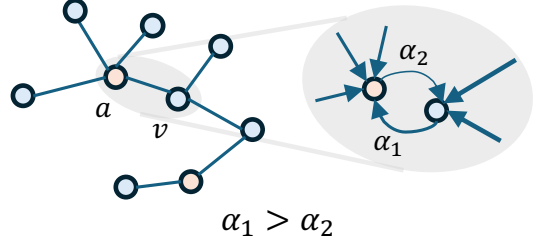


Figure 3: Symmetric attention: node a gives high attention to node v (α_1), while v gives lower attention to a (α_2). The final symmetric weight is $\min(\alpha_1, \alpha_2) = \alpha_2$.

Enhanced Message Passing. The computed attention weights modulate the message passing in the GNN architecture. For a node v , the attention-weighted messages from its neighbors are aggregated as follows:

$$\mathbf{h}_v^\ell = M(\{\tilde{A}^{\ell-1}[u, v] \cdot \mathbf{h}_u^{\ell-1} \mid u \in \mathcal{N}(v)\}) \quad (8)$$

where $M(\cdot)$ combines the weighted messages using sum aggregation followed by ReLU activation.

Unlike NSGT (Qiao & Pang, 2024), which removes edges based on a binary homophily threshold, NSAW introduces a learnable and continuous mechanism to modulate edge importance. This makes NSAW both more flexible and more robust: it preserves useful structure while suppressing anomalous influences during message passing. Instead of pruning edges, NSAW learns to softly attenuate misleading connections, especially those from anomalous nodes, making it more general and data-driven for cross-domain GAD scenarios.

4.4 Class-aware Regularization

Class imbalance poses a major challenge in GAD. As shown in Table 5, the abnormality rate of all datasets we’re using is less than 7%. Our class-aware regularization enhances the model’s ability to handle class imbalance during the source training phase by applying stronger regularization to the minority class (i.e., anomalous nodes). When minimizing similarity between non-connected nodes, it assigns higher weights to anomalous nodes, ensuring they maintain their distinctive patterns rather than being overshadowed by the majority class representations. Here’s the mathematical formulation:

$$\mathcal{L}_s = \sum_{i \in \mathcal{V}} \left(\frac{\sum_{j \in \mathcal{N}_i^-} w_j \cdot \text{sim}(\mathbf{h}_i, \mathbf{h}_j)}{|\mathcal{N}_i^-|} \right) \quad (9)$$

where \mathcal{N}_i^- represents the set of nodes not connected to node i , $\text{sim}(h_i, h_j)$ is the cosine similarity between node embeddings, and w_j is the class-based weight defined as:

$$w_j = \begin{cases} \alpha & \text{if node } j \text{ is anomalous} \\ 1 & \text{otherwise} \end{cases} \quad (10)$$

where $\alpha > 1$ is the weighting factor for anomalous nodes. We set α inversely proportional to the class ratio. This regularization can only be applied during source training where we have access to node labels, while target training relies solely on the homophily-based loss due to its unsupervised nature.

4.5 Test-Time Training

We can now describe our framework for anomaly detection with domain adaptation (GADT3), summarized in Figure 2. Following the test-time training paradigm, GADT3 is based on two tasks, the *main (supervised) task* and the *auxiliary self-supervised learning (SSL) task*. In the training phase, GADT3 is trained based on both tasks using a joint loss minimized based on the source dataset. Subsequently, the test-time training phase leverages only the SSL task using the target dataset, which is assumed to be unlabeled.

Training phase. During training, GADT3 applies the source graph \mathcal{G}_s to learn an encoder θ_s , a decoder θ_m , and a predictor θ_{pred} by minimizing a loss function $\mathcal{L}_{\text{train}}$ that is a combination of a supervised loss \mathcal{L}_{sup} , for which we apply cross-entropy, and a self-supervised loss $\mathcal{L}_{\text{self}}$, for which we apply the homophily-based affinity scores (see Sec. 4.2):

$$\mathcal{L}_{\text{sup}} = \sum_{v \in \mathcal{V}_s} y_v \log \hat{y}_v + \lambda_s \mathcal{L}_s \quad (11)$$

$$\mathcal{L}_{\text{self}} = - \sum_{v \in \mathcal{V}} \left(\frac{1}{|N(v)|} \sum_{u \in N(v)} \text{sim}(\mathbf{h}_v^L, \mathbf{h}_u^L) \right) + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} \quad (12)$$

$$\mathcal{L}_{\text{train}} = \sum_{v \in \mathcal{V}_s} \left(\mathcal{L}_{\text{sup}}(v, y; \theta_m, \theta_s, \theta_{\text{pred}}) + \lambda \mathcal{L}_{\text{self}}(v; \theta_m, \theta_s) \right) \quad (13)$$

where \hat{y} is the predicted label (normal or anomaly) and λ , λ_s , and λ_{reg} are weight hyperparameters.

We describe the source encoder θ_s (projection) and decoder θ_m (GNN) in Section 4.1. The predictor θ_{pred} is a multi-layer perception that maps node embeddings to label predictions.

Test-time training phase. During the test-time training (TTT) phase, we apply the target graph \mathcal{G}_t to optimize a target encoder θ_t by minimizing the TTT loss \mathcal{L}_{TTT} :

$$\mathcal{L}_{\text{TTT}} = \sum_{v \in \mathcal{V}_t} \mathcal{L}_{\text{self}}(v; \theta_t, \theta_m) \quad (14)$$

where $\mathcal{L}_{\text{self}}$ is defined in Eqn. 12 and the GNN decoder θ_m is re-used from the training phase to embed target nodes.

Inference: Anomaly detection in the target domain is performed using the adapted encoder θ_t , which is optimized during test-time training (TTT) based solely on the target graph. The decoder θ_m , trained on the source domain, is reused without modification. The encoder θ_t maps target node features into a shared embedding space, and θ_m further refines these representations using the target graph structure. We do not use the prediction head θ_{pred} at inference time. Instead, we compute homophily scores in the embedding space generated by θ_m , rank the nodes based on these scores, and directly evaluate performance using AUROC and AUPRC.

Although θ_{pred} is not applied during inference, we observe the embeddings learned by the adapted target encoder θ_t remain well-aligned with the source decision boundary learned during training. This is shown in

Figure 4, where we show that source and target embeddings share similar separation patterns, indicating that the shared representation space preserves discriminative structure even without explicit prediction. While θ_{pred} is trained using source labels, it can technically be reused during inference. However, GADT3 relies on homophily-based node ranking. See Appendix A.6 for an alternative approach.

Stopping criteria. We propose an adaptive early-stopping strategy based on distance ratios between source and target domains. Our method computes the distance between target features and the average embeddings of two distinct source classes—normal and anomalous. The ratio of maximum to minimum distance serves as the stopping criterion, where a higher ratio indicates better domain adaptation as it reflects stronger alignment between target samples and their corresponding source class features. The training stops when no improvement in this ratio is observed for a predefined number of epochs. Our ratio-based approach measures relative alignment between target samples and source class distributions, rather than just overall domain similarity. See Appendix A.1 for details.

5 Experiments

We evaluate the proposed approach (GADT3) against multiple baselines across six graph datasets, including methods from both cross-domain adaptation and anomaly detection. Our analysis is complemented with ablation studies on key components of the model, such as source training, NSAW, and class-aware regularization. To assess scalability, we further test GADT3 on three large-scale graphs. Additionally, we investigate the model’s sensitivity to the homophily assumption by synthetically varying homophily levels to identify the threshold at which performance begins to degrade.

Code availability: Our source code is available for reproducibility purposes as an anonymous repository.¹

5.1 Experimental Setup

Datasets. We apply six datasets from diverse domains such as online shopping reviews, including Amazon (AMZ) (McAuley & Leskovec, 2013), YelpChi (Rayana & Akoglu, 2015), YelpHotel (HTL), and YelpRes (RES) (Ding et al., 2021), and social networks, including Reddit (RDT) (Kumar et al., 2018) and Facebook (FB) (Leskovec & McAuley, 2012). For our cross-domain analysis, we examine scenarios where feature spaces are homogeneous (same features) and heterogeneous (different features) across domains. To demonstrate the scalability of our model, we additionally report results on three large-scale graphs: Amazon-all (AMZ-all), YelpChi-all (YC-all) (Hamilton et al., 2017), and T-Finance (TF) (Tang et al., 2022) in Table 3. A detailed description of each dataset is provided in Appendix A.2.

Baselines and Evaluation Criteria. We evaluate our method against state-of-the-art baselines both for cross-domain domain adaptation and (single-graph) unsupervised anomaly detection. Cross-domain baselines include GRADE (Wu et al., 2023), AdaGCN (Dai et al., 2022), UDA-GCN (Wu et al., 2020), and ACT (Wang et al., 2023b). Test-time training baselines include TENT (Wang et al., 2020), GraphCL (You et al., 2020), and GTrans (Jin et al., 2022). As GAD baselines, we consider self-supervised learning-based methods (CoLA (Liu et al., 2021a), and SL-GAD (Zheng et al., 2021)), reconstruction-based methods (DOMINANT (Ding et al., 2019) and ComGA (Luo et al., 2022)), and a local affinity-based method (TAM (Qiao & Pang, 2024)).

We utilize two metrics to evaluate anomaly detection models: the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC). Higher values of AUROC and AUPRC indicate superior model performance. For all experiments, we report average results for 10 repetitions.

Implementation details. We developed GADT3 using Python 3.11.9 and PyTorch 2.1.0. All the experiments were conducted on an NVIDIA A40 GPU with 48GB. The core model is a 2-layer GraphSAGE GNN with weight parameters optimized via the Adam optimizer (Kingma & Ba, 2014) with a dropout rate of 0.7. The source model underwent training for 100 epochs with early stopping employed to determine the target epochs. Both the source and target models were trained with a learning rate of 0.001. We set $\lambda = 0.001$,

¹<https://anonymous.4open.science/r/GADT3-Algorithm-BE45/>

Table 1: Cross-domain (Source \rightarrow Target) node-level graph anomaly detection performance of GADT3 compared to state-of-the-art baselines. We have shown the results for domain pairs with both homogeneous and heterogeneous feature spaces. Results show GADT3 significantly outperforms baselines in most settings.

Metric (%)	Type	Method	Heterogeneous Features						Homogeneous Features		Avg.
			AMZ \rightarrow RDT	AMZ \rightarrow FB	RDT \rightarrow AMZ	RDT \rightarrow FB	FB \rightarrow AMZ	FB \rightarrow RDT	HTL \rightarrow RES	RES \rightarrow HTL	
AUROC	CDA	GRADE	61.80 \pm 0.8	84.96 \pm 0.7	70.38 \pm 1.2	85.53 \pm 0.6	72.06 \pm 0.9	61.27 \pm 1.0	75.51 \pm 0.6	73.80 \pm 0.5	73.16 \pm 0.4
		AdaGCN	58.86 \pm 0.7	72.35 \pm 0.8	70.43 \pm 0.6	75.89 \pm 1.1	64.86 \pm 1.3	61.62 \pm 0.9	79.50 \pm 0.7	81.18 \pm 0.5	70.59 \pm 0.6
		UDA-GCN	60.37 \pm 1.0	80.80 \pm 0.9	67.36 \pm 1.2	84.98 \pm 0.6	67.55 \pm 1.0	61.95 \pm 0.8	85.36 \pm 0.9	78.24 \pm 0.6	73.33 \pm 0.7
		ACT	59.24 \pm 1.3	81.97 \pm 0.5	71.16 \pm 0.6	75.74 \pm 1.1	71.98 \pm 1.0	60.52 \pm 1.2	89.20 \pm 0.4	80.40 \pm 0.8	73.78 \pm 0.6
	TTT	TENT	59.01 \pm 0.9	76.15 \pm 0.8	71.51 \pm 0.5	83.43 \pm 0.7	66.88 \pm 1.2	58.20 \pm 0.9	75.11 \pm 1.0	79.27 \pm 0.5	71.20 \pm 0.5
		GraphCL	60.37 \pm 0.7	88.46 \pm 0.6	68.71 \pm 0.8	88.81 \pm 0.7	61.91 \pm 1.0	61.64 \pm 0.9	79.05 \pm 0.5	74.20 \pm 0.6	72.89 \pm 0.7
		GTrans	60.64 \pm 0.6	89.35 \pm 0.5	73.93 \pm 0.9	90.18 \pm 0.4	68.03 \pm 0.7	61.90 \pm 0.8	79.86 \pm 0.7	83.49 \pm 0.5	75.92 \pm 0.6
	Ours	GADT3	61.95 \pm 0.5	91.03 \pm 0.2	79.87 \pm 0.3	94.84 \pm 0.3	82.57 \pm 0.3	61.93 \pm 0.7	94.14 \pm 0.1	90.71 \pm 0.1	82.13 \pm 0.3
	CDA	GRADE	4.92 \pm 0.2	18.44 \pm 0.3	11.87 \pm 0.5	16.10 \pm 0.4	13.00 \pm 0.3	4.34 \pm 0.2	18.81 \pm 0.6	22.53 \pm 0.4	13.75 \pm 0.3
		AdaGCN	4.51 \pm 0.3	9.04 \pm 0.2	12.15 \pm 0.4	5.72 \pm 0.5	10.23 \pm 0.4	4.64 \pm 0.3	22.15 \pm 0.4	35.15 \pm 0.5	12.95 \pm 0.3
		UDA-GCN	4.65 \pm 0.2	6.73 \pm 0.3	10.85 \pm 0.3	22.04 \pm 0.6	10.89 \pm 0.4	5.55 \pm 0.3	26.02 \pm 0.5	27.31 \pm 0.4	14.26 \pm 0.3
		ACT	4.22 \pm 0.4	8.16 \pm 0.3	13.20 \pm 0.5	26.52 \pm 0.4	12.26 \pm 0.3	4.88 \pm 0.2	33.00 \pm 0.3	28.70 \pm 0.5	16.37 \pm 0.4
AUPRC	TTT	TENT	4.60 \pm 0.3	11.69 \pm 0.4	12.49 \pm 0.3	18.81 \pm 0.5	12.32 \pm 0.4	4.79 \pm 0.3	13.12 \pm 0.3	28.31 \pm 0.4	13.27 \pm 0.4
		GraphCL	4.69 \pm 0.2	21.09 \pm 0.3	14.36 \pm 0.4	30.86 \pm 0.3	11.12 \pm 0.5	5.44 \pm 0.3	21.27 \pm 0.4	25.55 \pm 0.3	16.80 \pm 0.3
		GTrans	4.78 \pm 0.3	25.95 \pm 0.4	17.02 \pm 0.5	29.11 \pm 0.2	11.91 \pm 0.4	5.15 \pm 0.3	21.03 \pm 0.4	36.20 \pm 0.3	18.89 \pm 0.4
	Ours	GADT3	5.19 \pm 0.1	27.06 \pm 0.1	19.10 \pm 0.4	34.76 \pm 0.3	25.73 \pm 0.4	6.12 \pm 0.2	39.75 \pm 0.1	42.90 \pm 0.3	25.08 \pm 0.3

$\lambda_{\text{reg}} = 0.1$, $\lambda_s = 0.001$, $p = 40$, and $\alpha = 20$. Baselines were trained using the hyperparameters recommended in their respective papers.

5.2 Cross-domain Graph Anomaly Detection

We present the results comparing our approach against various domain adaptation baselines in Table 1. We use different datasets as source and target domains. We incorporated a projection head into all baseline models to address feature sets’ mismatches across source and target domain datasets. This addition enables fair comparison by allowing each method to handle the discrepancy in feature spaces.

Our proposed method (GADT3) outperformed the baselines in most scenarios. GADT3 achieved the highest AUROC and AUPRC scores in 5 out of 6 tasks. For instance, in terms of AUROC, in the Reddit \rightarrow Amazon task, GADT3 surpassed the next best method (GTrans) by 5.9%. In the Reddit \rightarrow Facebook task, GADT3 outperformed the closest competitor (GTrans) by 4.7%. The only task where GADT3 did not lead in terms of AUROC and AUPRC was Facebook \rightarrow Reddit, where it performed competitively but slightly below UDA-GCN. On average, our method outperformed the second-best baseline by at least 6.2 points in both AUROC and AUPRC. In Appendix A.5, we provide further insights into the performance of GADT3. We have found that high target performance is often correlated with high source homophily. This can be explained by our use of a homophily-based self-supervised loss for test-time training.

5.3 Representation Embedding

The effectiveness of our TTT approach is demonstrated in Figure 4, which presents the two-dimensional T-SNE embeddings of the GNN outputs for both source and target models. On the left side, we observe a clear separation between normal and anomalous samples in the source domain (Amazon), indicating that the source model has effectively learned to distinguish between classes. The right figure shows the transferability of our approach to the target domain (Facebook). Despite the inherent differences between source and target datasets, the adapted embeddings in the target domain preserve meaningful structure, demonstrating that our TTT strategy successfully refines the target representations using the pre-trained source model. We provide additional visualizations of source and target embeddings for other datasets in Appendix A.7.

5.4 Early Stopping

This section evaluates our early-stopping strategy across different domain adaptation scenarios. Figure 5 illustrates how the distance ratio evolves together with the Area Under the Precision-Recall Curve (AUPRC)

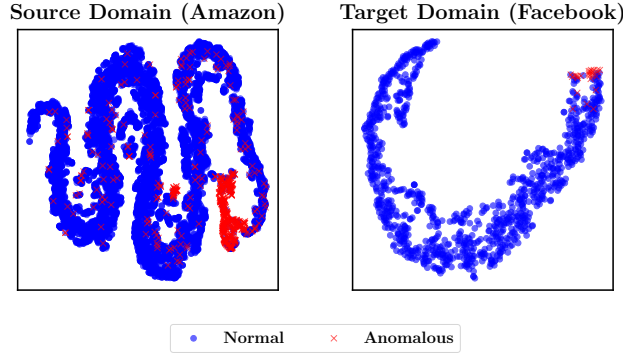


Figure 4: T-SNE embeddings of node representations for source domain (Amazon) on the left and target domain (Facebook) on the right. The source model clearly separates normal/anomalous nodes. After adaptation, the pre-trained model maintains this separation in target domain, indicating successful transfer.

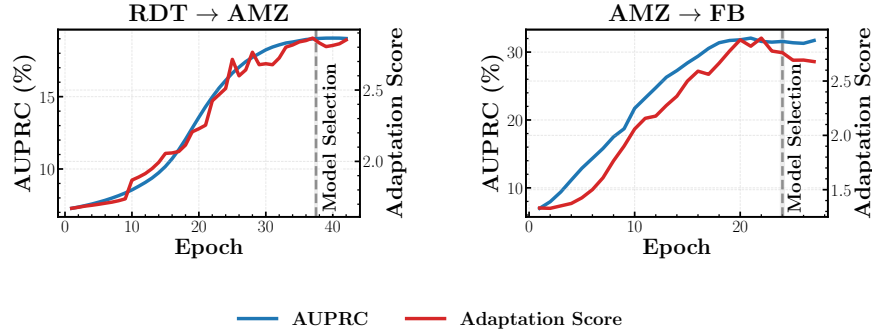


Figure 5: Relationship between target anomaly detection accuracy (AUPRC) and our ratio-based adaptation score for Reddit→Amazon and Amazon→Facebook. We use the adaptation score as a label-free early stopping criterion, which is triggered at epochs 37 (left) and 24 (right).

during training. We marked the specific epoch when the model was selected. The results show that the ratio effectively captures the convergence of domain adaptation without the need for labeled validation data.

6 Ablation Studies

We perform four ablation studies to evaluate different aspects of GADT3. The results support the design of our model and demonstrate the relevance of each of its components.

GAD Results (Source-free). We evaluate a single-domain version of GADT3 against six state-of-the-art GAD baselines using four datasets. Our goal is to isolate the impact of GADT3’s self-supervised learning from its cross-domain adaptation. The results are shown in Table 2. We follow the same protocol and use the same datasets as Qiao & Pang (2024), so we report their results for the baselines. GADT3 outperforms baselines even in the source-free setting, as the NSAW edge weighting effectively leverages structural context to enhance anomaly separation without relying on source data.

NSAW as a learnable and flexible edge weighting approach. We proposed NSAW as a learnable edge-weighting mechanism to suppress the adverse effect of non-homophily edges in anomaly detection. Table 2 compares our method against TAM (Qiao & Pang, 2024) (using NSGT). The results show that GADT3 outperforms TAM using most datasets. This is evidence that NSAW is a more flexible alternative to NSGT by allowing GADT3 to learn continuous attention weights. To validate its effectiveness, we compare GADT3 with and without NSAW under our cross-domain tasks. Introducing NSAW improves AUROC from 90.83%

Table 2: GADT3, trained on target only, outperforms specialized baselines on most datasets.

Metric (%)	Method	AMZ	RDT	FB	YC	Avg.
AUROC	CoLA	58.98	60.28	84.34	46.36	62.49
	SL-GAD	59.37	56.77	79.36	33.12	57.16
	DOMINANT	59.96	55.55	56.77	41.33	53.40
	ComGA	58.95	54.53	60.55	43.91	54.48
	TAM	70.64	60.23	91.44	56.43	69.69
	GADT3	84.76	61.54	91.74	48.67	71.68
AUPRC	CoLA	6.77	4.49	21.06	4.48	9.20
	SL-GAD	6.34	4.06	13.16	3.50	6.76
	DOMINANT	14.24	3.56	3.14	3.95	6.22
	ComGA	11.53	3.74	3.54	4.23	5.76
	TAM	26.34	4.46	22.33	7.78	15.23
	GADT3	27.53	6.01	26.13	6.14	16.45

Table 3: AUROC and AUPRC (%) on large-scale graphs. **Bold** indicates best.

Metric (%)	Method	AMZ-all	YC-all	TF
AUROC	CoLA	26.14	48.01	48.29
	SL-GAD	27.28	55.51	46.48
	DOMINANT	69.37	53.90	53.80
	ComGA	71.54	53.52	55.42
	TAM	84.76	58.18	61.75
	GADT3	86.23	55.31	62.91
AUPRC	CoLA	5.16	13.61	4.10
	SL-GAD	4.44	17.11	3.86
	DOMINANT	10.15	16.38	4.74
	ComGA	18.54	16.58	4.81
	TAM	43.46	18.86	5.47
	GADT3	44.98	17.77	6.23

to 91.04% on AMZ→FB and from 82.97% to 86.11% on FB→AMZ. These gains confirm that learnable, structure-preserving edge weighting is more effective than static edge removal strategies.

Performance on Large-Scale Graphs. We evaluate GADT3 on three large-scale datasets and compare it against strong GAD baselines in Table 3. GADT3 outperforms all baselines on Amazon-all and T-Finance, and performs competitively on YelpChi-all.

Robustness Under Low Homophily Settings.

Our self-supervised loss for anomaly detection is based on homophily, *i.e.*, the tendency of similar nodes to connect. To assess the generalization of GADT3 in settings where this assumption weakens and beyond homophilic graphs, we synthetically rewired a portion of edges to reduce homophily in the AMZ and RDT datasets. As shown in Table 4, GADT3 remained robust when the target domain’s average homophily score was ≥ 0.5 , with only gradual degradation in performance below that threshold. These results show the robustness of our method while acknowledging its reliance on moderate levels of homophily.

Table 4: AUROC (%) under decreasing target-domain homophily.

Homophily	AMZ→RDT	RDT→AMZ
0.9	62.3	81.3
0.7	62.7	79.6
0.5	60.9	79.1
0.3	58.3	63.1
0.1	58.8	63.7

7 Conclusion

In this paper, we have introduced GADT3, a novel TTT framework for GAD when the testing data is out-of-distribution and from heterogeneous application domains. Specifically, our approach addresses scenarios where the distributions and feature spaces differ between the source and target datasets using dataset-specific encoders. GADT3 combines the advantages of TTT with multiple innovations focused on graph anomaly detection. For instance, by leveraging the distinctive distribution pattern of node homophily between normal and anomalous nodes, we have presented a tailored SSL task along with a robust attention-based edge-weighting mechanism (NSAW) to enhance the generalization of the learned representations. Moreover, GADT3 performs model election using an distance ratio-based early-stopping criterion. Our experiments have illustrated that GADT3 outperforms state-of-the-art graph domain adaptation baselines.

Limitations: One current limitation of our approach is its reliance on the assumption that normal nodes exhibit higher homophily compared to anomalous ones. While this pattern holds across many real-world datasets, relaxing the homophily assumption remains an important direction for future work to further enhance generalization to highly heterophilic graphs. Moreover, we will study theoretical guarantees for cross-domain GAD based on the similarity between source and target datasets in future works.

References

- Leman Akoglu, Mary McGlohon, and Christos Faloutsos. Oddball: Spotting anomalies in weighted graphs. In *Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part II 14*, pp. 410–421. Springer, 2010.
- Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery*, 29:626–688, 2015.
- Guanzi Chen, Jiyang Zhang, Xi Xiao, and Yang Li. Graphhta: Test time adaptation on graph neural networks. *arXiv preprint arXiv:2208.09126*, 2022.
- Nan Chen, Zemin Liu, Bryan Hooi, Bingsheng He, Rizal Fathony, Jun Hu, and Jia Chen. Consistency training with learnable data augmentation for graph anomaly detection with limited supervision. In *The Twelfth International Conference on Learning Representations*, 2024.
- Qingchao Chen and Yang Liu. Structure-aware feature fusion for unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10567–10574, 2020.
- Quanyu Dai, Xiao-Ming Wu, Jiaren Xiao, Xiao Shen, and Dan Wang. Graph transfer learning via adversarial domain adaptation with graph convolution. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4908–4922, 2022.
- Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM international conference on data mining*, pp. 594–602. SIAM, 2019.
- Kaize Ding, Kai Shu, Xuan Shan, Jundong Li, and Huan Liu. Cross-domain graph anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2406–2415, 2021.
- Zhengming Ding, Sheng Li, Ming Shao, and Yun Fu. Graph adaptive knowledge transfer for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 37–52, 2018.
- Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.
- Tianjin Huang, Yulong Pei, Vlado Menkovski, and Mykola Pechenizkiy. Hop-count based self-supervised anomaly detection on attributed networks. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 225–241. Springer, 2022.
- Jianguo Jiang, Jiuming Chen, Tianbo Gu, Kim-Kwang Raymond Choo, Chao Liu, Min Yu, Weiqing Huang, and Prasant Mohapatra. Anomaly detection with graph convolutional networks for insider threat and fraud detection. In *MILCOM 2019-2019 IEEE Military Communications Conference (MILCOM)*. IEEE, 2019.
- Wei Jin, Tong Zhao, Jiayuan Ding, Yozen Liu, Jiliang Tang, and Neil Shah. Empowering graph representation learning with test-time graph transformation. *arXiv preprint arXiv:2210.03561*, 2022.
- Ilhan Firat Kilincer, Fatih Ertam, and Abdulkadir Sengur. Machine learning methods for cyber security intrusion detection: Datasets and comparative study. *Computer Networks*, 188:107840, 2021.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. Community interaction and conflict on the web. In *Proceedings of the 2018 world wide web conference*, pp. 933–943, 2018.
- Aleksandar Lazarevic, Levent Ertoz, Vipin Kumar, Aysel Ozgur, and Jaideep Srivastava. A comparative study of anomaly detection schemes in network intrusion detection. In *Proceedings of the 2003 SIAM international conference on data mining*, pp. 25–36. SIAM, 2003.

- Jure Leskovec and Julian McAuley. Learning to discover social circles in ego networks. *Advances in neural information processing systems*, 25, 2012.
- Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. Out-of-distribution generalization on graphs: A survey. *arXiv preprint arXiv:2202.07987*, 2022.
- Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.
- Yixin Liu, Zhao Li, Shirui Pan, Chen Gong, Chuan Zhou, and George Karypis. Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE transactions on neural networks and learning systems*, 33(6):2378–2392, 2021a.
- Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34:21808–21820, 2021b.
- Xuexiong Luo, Jia Wu, Amin Beheshti, Jian Yang, Xiankun Zhang, Yuan Wang, and Shan Xue. Comga: Community-aware attributed graph anomaly detection. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 657–665, 2022.
- Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z Sheng, Hui Xiong, and Leman Akoglu. A comprehensive survey on graph anomaly detection with deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12012–12038, 2021.
- Xiaoxiao Ma, Ruikun Li, Fanzhen Liu, Kaize Ding, Jian Yang, and Jia Wu. Graph anomaly detection with few labels: A data-centric approach. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2153–2164, 2024.
- Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 897–908, 2013.
- Hezhe Qiao and Guansong Pang. Truncated affinity maximization: One-class homophily modeling for graph anomaly detection. *Advances in Neural Information Processing Systems*, 36, 2024.
- Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*, pp. 985–994, 2015.
- Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- Boshen Shi, Yongqing Wang, Fangda Guo, Bingbing Xu, Huawei Shen, and Xueqi Cheng. Graph domain adaptation: Challenges, progress and prospects. *arXiv preprint arXiv:2402.00904*, 2024.
- Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.
- Jianheng Tang, Jiajin Li, Ziqi Gao, and Jia Li. Rethinking graph neural networks for anomaly detection. In *International conference on machine learning*, pp. 21076–21089. PMLR, 2022.
- Chih-Fong Tsai, Yu-Feng Hsu, Chia-Ying Lin, and Wei-Yang Lin. Intrusion detection by machine learning: A review. *expert systems with applications*, 36(10):11994–12000, 2009.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

- M Venkatesan and P Prabhavathy. Graph based unsupervised learning methods for edge and node anomaly detection in social network. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pp. 1–5. IEEE, 2019.
- Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.
- Huan Wang, Guoquan Liu, and Po Hu. Tdan: transferable domain adversarial network for link prediction in heterogeneous social networks. *ACM Transactions on Knowledge Discovery from Data*, 18(1):1–22, 2023a.
- Qizhou Wang, Guansong Pang, Mahsa Salehi, Wray Buntine, and Christopher Leckie. Cross-domain graph anomaly detection via anomaly-aware contrastive alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4676–4684, 2023b.
- Xuhong Wang, Baihong Jin, Ying Du, Ping Cui, Yingshui Tan, and Yupu Yang. One-class graph neural networks for anomaly detection in attributed networks. *Neural computing and applications*, 33:12073–12085, 2021.
- Yiqi Wang, Chaozhuo Li, Wei Jin, Rui Li, Jianan Zhao, Jiliang Tang, and Xing Xie. Test-time training for graph neural networks. *arXiv preprint arXiv:2210.08813*, 2022.
- Jun Wu, Jingrui He, and Elizabeth Ainsworth. Non-iid transfer learning on graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10342–10350, 2023.
- Man Wu, Shirui Pan, Chuan Zhou, Xiaojun Chang, and Xingquan Zhu. Unsupervised domain adaptive graph convolutional networks. In *Proceedings of the web conference 2020*, pp. 1457–1467, 2020.
- Man Wu, Xin Zheng, Qin Zhang, Xiao Shen, Xiong Luo, Xingquan Zhu, and Shirui Pan. Graph learning under distribution shifts: A comprehensive survey on domain adaptation, out-of-distribution, and continual learning. *arXiv preprint arXiv:2402.16374*, 2024.
- Nan Yin, Li Shen, Mengzhu Wang, Long Lan, Zeyu Ma, Chong Chen, Xian-Sheng Hua, and Xiao Luo. Coco: A coupled contrastive framework for unsupervised domain adaptive graph classification. In *International Conference on Machine Learning*, pp. 40040–40053. PMLR, 2023.
- Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823, 2020.
- Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. Graph domain adaptation via theory-grounded spectral regularization. In *The eleventh international conference on learning representations*, 2023.
- Ge Zhang, Zhao Li, Jiaming Huang, Jia Wu, Chuan Zhou, Jian Yang, and Jianliang Gao. efraudcom: An e-commerce fraud detection system via competitive graph neural networks. *ACM Transactions on Information Systems (TOIS)*, 40(3):1–29, 2022.
- Jiaxin Zhang, Yiqi Wang, Xihong Yang, and En Zhu. A fully test-time training framework for semi-supervised node classification on out-of-distribution graphs. *ACM Transactions on Knowledge Discovery from Data*, 2024.
- Li Zheng, Zhenpeng Li, Jian Li, Zhao Li, and Jun Gao. Addgraph: Anomaly detection in dynamic graph using attention-based temporal gen. In *IJCAI*, volume 3, pp. 7, 2019.
- Yu Zheng, Ming Jin, Yixin Liu, Lianhua Chi, Khoa T Phan, and Yi-Ping Phoebe Chen. Generative and contrastive self-supervised learning for graph anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12), 2021.