Data Engineering and Applications

Research Paper

Anomaly Detection in Incomplete BIM Models Based on Contrastive Learning

Damian Wisniewski[1, *] and Aleksandra Wojcik[2]

[1] Faculty of Computer Science and Information Technology, Warsaw University of Technology, 00-665 Warsaw, Poland
[2] Faculty of Mathematics and Information Science, Gdansk University of Technology, 80-233 Gdansk, Poland
*Corresponding author: damian.w@pw.edu.pl

Abstract. Building Information Modeling (BIM) has become a cornerstone in digital construction, yet the prevalence of incomplete data poses significant challenges for reliable anomaly detection and quality assurance. This study aims to develop a robust anomaly detection framework tailored for BIM environments characterized by missing, inconsistent, or heterogeneous information. A deep contrastive learning architecture is proposed, integrating graph-based representation learning with explicit missingness modeling and data augmentation strategies. The model constructs positive and negative pairs from augmented BIM samples, enabling the network to learn invariant and discriminative features despite varying degrees of data incompleteness. Experimental evaluation is conducted on a comprehensive BIM dataset comprising both authentic and systematically simulated missing data. Results indicate that the proposed framework achieves superior precision, recall, F1-score, and AUC compared to traditional machine learning and deep learning baselines, particularly under conditions of high or structured missingness. Ablation studies demonstrate the critical role of each architectural component, and robustness analysis confirms the method's stability when facing extreme data loss. These findings underscore the practical value of the approach for real-world BIM quality control, as it enables accurate and reliable detection of semantic and topological anomalies without requiring full data integrity. The research concludes by outlining prospects for deploying the framework at scale, integrating with industrial BIM platforms, and addressing more complex anomaly types in future work.

Keywords: *Contrastive Learning, Anomaly Detection, Graph Neural Networks, Building Information Modeling*

Received on 29 September 2023, Accepted on 5 November 2024, Published on 13 Jan2025

Introduction

Building Information Modeling (BIM) has emerged as a transformative technology in the architecture, engineering, and construction (AEC) industries, offering a comprehensive digital representation of physical and functional characteristics throughout the lifecycle of buildings and infrastructure [1]. By enabling seamless data integration, visualization, and collaborative workflows, BIM significantly enhances decision-making, operational efficiency, and project outcomes [2]. As BIM adoption accelerates globally, the quality and completeness of BIM data have become critical determinants of its utility in downstream applications, such as automated construction monitoring, facility management, and predictive maintenance [3]. However, BIM data is frequently subject to incompleteness due to a variety of factors, including human error during data entry, intermittent sensor failures, heterogeneous data sources, and the complex, dynamic nature of construction sites [4]. Data gaps, inconsistencies, and missing attributes not only undermine the effectiveness of BIM-based analytics but also pose significant challenges for the reliability and scalability of intelligent automation in the built environment [5,6].

Anomaly detection in BIM models plays a pivotal role in ensuring construction quality, safety, and compliance by identifying irregularities such as design deviations, structural inconsistencies, and abnormal operational patterns [7]. Traditional anomaly detection approaches, including rule-based systems, statistical models, and

conventional machine learning algorithms, typically assume the availability of complete and high-quality datasets [8]. In real-world scenarios, however, the presence of missing or corrupted data severely hampers the performance of these methods, leading to increased false alarms, undetected anomalies, and suboptimal decision support. Existing solutions for handling incomplete data—such as data imputation, robust statistical estimation, or ignoring missing values—often introduce additional bias or information loss, ultimately limiting model generalizability and interpretability. Furthermore, deep learning-based anomaly detection techniques, while demonstrating superior performance in various computer vision and time-series tasks, are also constrained by their reliance on large-scale, fully annotated, and complete datasets. The inherent sparsity and heterogeneity of BIM data, compounded by the high dimensionality and complex interdependencies among building elements, necessitate more robust and adaptive approaches capable of learning meaningful representations from partial and noisy inputs.

To address these challenges, this paper proposes a novel deep contrastive learning framework specifically designed for anomaly detection in BIM models under conditions of data incompleteness. By leveraging the power of contrastive representation learning, our approach enables the extraction of discriminative and robust features from partially observed BIM data without relying on extensive manual annotation or data imputation. The main contributions of this work are threefold: 1) Formulate the problem of BIM anomaly detection with incomplete data and analyze the limitations of conventional methods; 2) Develop a deep contrastive learning architecture that effectively captures semantic similarities and discrepancies between incomplete BIM samples, thereby enhancing detection performance; and 3) Conduct comprehensive experiments on both synthetic and real-world BIM datasets, demonstrating the superiority and robustness of the proposed approach compared to state-of-the-art baselines. The remainder of this paper is structured as follows: Section 2 reviews related work on BIM anomaly detection, data incompleteness management, and contrastive learning; Section 3 details the proposed methodology, including problem formulation, model architecture, and optimization strategies; Section 4 describes the experimental setup, datasets, and evaluation metrics; Section 5 presents results and in-depth discussion; and Section 6 concludes with a summary of findings and future research directions.

Related Work

BIM Anomaly Detection Techniques

The detection of anomalies in Building Information Modeling (BIM) has attracted significant research interest, given its implications for construction quality, schedule integrity, and operational safety. Early approaches primarily relied on rule-based systems, where domain experts encoded explicit logical rules or constraints to flag inconsistencies, deviations, or missing elements in BIM datasets [9]. While effective to some extent, these systems are inherently limited by their inability to generalize beyond predefined scenarios and are highly sensitive to data noise and incompleteness.

Subsequent advances introduced traditional machine learning methods, such as clustering, support vector machines, and decision trees, to capture more complex patterns and automate anomaly detection within BIM environments [10]. However, these algorithms typically assume that the input data is complete and well-structured, an assumption that rarely holds in real-world BIM applications. Missing values, heterogeneity in attribute formats, and partial observations often lead to performance degradation, increased false positives, and reduced reliability [11].

In recent years, deep learning-based techniques have demonstrated remarkable capability in learning hierarchical representations and identifying subtle irregularities in high-dimensional BIM data [12]. Convolutional neural networks (CNNs), graph neural networks (GNNs), and autoencoders have been adapted for tasks ranging from defect classification to semantic segmentation and model verification. Nonetheless, the effectiveness of these data-driven approaches is predicated on the availability of large-scale, complete, and accurately annotated BIM datasets—a condition that is seldom met in practice. As a result, deep learning models frequently struggle with robustness and generalizability when exposed to incomplete or corrupted BIM data [13]. This recurring challenge motivates the exploration of more adaptive and resilient anomaly detection frameworks, especially those capable of operating under data incompleteness.

DOI: 10.64972/dea.2025.v3i1.81

Data Incompleteness Handling

Handling data incompleteness is a foundational concern in data-centric engineering research. Traditional strategies such as mean or median imputation, k-nearest neighbor imputation, and matrix factorization have long been employed to fill missing values in tabular and structured datasets [14]. In the context of BIM, these methods are often inadequate due to the high dimensionality and intricate dependencies among building components, which can render simple imputation techniques both ineffective and potentially misleading.

More advanced solutions have emerged, including probabilistic graphical models, variational autoencoders, and generative adversarial networks, all of which seek to reconstruct missing information by leveraging correlations in the observed data [15]. Robust learning paradigms, which focus on designing loss functions and model architectures that are inherently insensitive to missing or noisy inputs, have also gained traction. Some recent works in the BIM domain have explored hybrid approaches that combine data completion with anomaly detection, using iterative refinement or semi-supervised learning to improve resilience against data gaps [16].

Despite these advances, the transferability and scalability of such methods to large, heterogeneous BIM datasets remain limited. Imputation-based techniques may introduce bias or obscure true anomalies, while robust learning frameworks often require careful tuning and extensive domain knowledge. Therefore, there is a growing consensus that anomaly detection in BIM should move beyond mere data completion and seek approaches that directly leverage the available, albeit incomplete, information. This perspective underpins the motivation for exploring contrastive learning as a more principled and flexible solution.

Contrastive Learning in Computer Vision

Contrastive learning has emerged as a powerful paradigm in representation learning, particularly within computer vision. The core idea is to learn an embedding space in which similar (positive) pairs are pulled together, while dissimilar (negative) pairs are pushed apart [17]. Pioneering frameworks such as SimCLR and MoCo have demonstrated the effectiveness of this approach in unsupervised and self-supervised settings, enabling models to learn rich and discriminative features from unlabeled and even incomplete data.

One of the key strengths of contrastive learning lies in its ability to exploit the inherent structure of the data, making it well-suited for scenarios where annotations are scarce or data is partially missing. By constructing positive and negative sample pairs, contrastive methods can extract robust semantic representations that generalize well, even when facing significant data incompleteness. Recent research has extended these techniques to domains such as 3D vision and graph-structured data, highlighting their versatility and adaptability.

Despite its promise, the application of contrastive learning to BIM anomaly detection under incomplete data remains largely unexplored. This study aims to bridge this gap by proposing a deep contrastive learning framework tailored to the unique challenges of BIM data, leveraging the strengths of contrastive paradigms to enhance anomaly detection under real-world data constraints.

Proposed Methodology

Problem Formulation

Building Information Modeling (BIM) provides a rich, structured digital representation of built assets, encapsulating geometric, semantic, and relational information. For a given project, let us formally denote the BIM model as a graph structure $\mathcal{G} - (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where $\mathcal{V}$ represents the set of building elements (nodes), $\mathcal{E}$ the set of topological or functional relationships (edges), and $\mathcal{A}$ the set of attributes associated with each node and edge. Each node $v_i \in \mathcal{V}$ is described by a feature vector $\mathbf{x}_i \in \mathbb{R}^d$, capturing properties such as type, material, and geometric parameters. The global BIM data matrix is then defined as $\mathbf{X} - [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times d}$, where $n - |\mathcal{V}|$ and d is the feature dimension.

In practice, BIM datasets are often incomplete, either due to missing attribute values, omitted elements, or corrupted relationship data. We introduce a binary mask matrix $\mathbf{M} \in \{0,1\}^{n \times d}$, where $m_{ij} - 1$ if the j-th attribute of node i is observed, and $m_{ij} - 0$ otherwise. The incomplete BIM data can thus be represented as $\overline{\mathbf{X}} - \mathbf{X} \odot \mathbf{M}$, where $\odot$ denotes the Hadamard (element-wise) product. This formalism generalizes across

DOI: 10.64972/dea.2025.v3i1.81

different missing patterns, including random missingness and systematic omissions caused by specific data acquisition failures [18].

The anomaly detection objective is to learn a decision function $f: (\overline{\mathbf{X}}, \mathcal{E}, \mathcal{A}) \rightarrow \mathbf{y}$, where $\mathbf{y} \in \{0,1\}^n$ is a binary vector indicating the presence ($y_i - 1$) or absence ($y_i - 0$) of anomalies in each building element. Anomalies may manifest as attribute outliers, structural inconsistencies, or unusual relational patterns, and their identification must be robust to varying degrees of data incompleteness. The problem can be summarized by the following equations:

(1) Incomplete BIM data representation:

$$\overline{\mathbf{X}} = \mathbf{X} \odot \mathbf{M} \quad \text{Eq. (1)}$$

(2) Anomaly detection target:

$$\mathbf{y} = f(\overline{\mathbf{X}}, \mathcal{E}, \mathcal{A}) \quad \text{Eq. (2)}$$

For example, consider a real-world scenario where sensor malfunctions lead to missing temperature readings on HVAC components, or as-built model updates omit recently installed structural elements. In such cases, traditional anomaly detection models trained on fully observed data may fail to flag irregularities, leading to overlooked risks or false alarms [19]. Therefore, robust methods are needed to directly process incomplete BIM data and accurately identify anomalous elements without requiring prior imputation or manual correction.

In the following section introduce a deep contrastive learning framework engineered to address this challenge. Our approach leverages the structural and semantic context within incomplete BIM representations to learn discriminative features that are resilient to missing information, thereby enabling reliable anomaly detection in practical, imperfect datasets [20,21].

Deep Contrastive Learning Architecture

The proposed deep contrastive learning architecture is tailored for robust anomaly detection in incomplete BIM datasets. The complete workflow is depicted in Figure 1, which illustrates the progression from data preprocessing through deep encoding and projection, up to the anomaly scoring output. This system is designed to fully leverage both the relational structure and attribute semantics inherent in BIM representations, thus significantly enhancing resilience to missing data.

As shown in Figure 1, the process begins with the input of the incomplete BIM graph, where each building element is represented by a node-attribute matrix together with its relational graph. During preprocessing, missing attributes are explicitly encoded with indicator variables. Standardization and categorical encoding ensure the data is compatible with the network input, while missingness is preserved as a feature for the model to learn from, rather than being imputed. To improve generalization across different missingness patterns, the data augmentation module generates multiple views for each building element. These augmentations include random attribute dropout, local topology perturbation, and the injection of synthetic noise, all of which mirror the types of incompleteness encountered in real-world BIM acquisition, maintenance, and transmission scenarios.

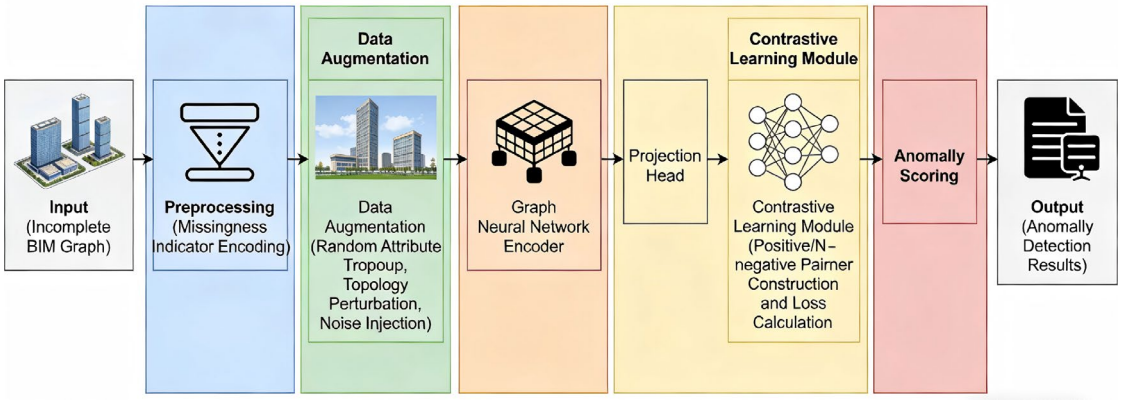

Figure 1. Overview of the proposed deep contrastive learning framework for anomaly detection in incomplete BIM data.

DOI: 10.64972/dea.2025.v3i1.81

A pivotal aspect of the framework is the construction of positive and negative sample pairs for contrastive learning. Positive pairs are formed by taking two different augmented views of the same building element, while negative pairs are sampled from distinct elements within a batch. This sample pairing strategy ensures that the encoder focuses on learning invariant semantic features that persist despite differing observed attributes or topological disturbances, as required for resilient anomaly detection.

The core of the architecture is a multi-layer graph neural network encoder. For each node v_i with its incomplete attribute vector $\overline{\mathbf{x}}_i$ and neighborhood $\mathcal{N}(v_i)$, the encoder computes a latent embedding as follows:

$$\mathbf{z}_i = \mathrm{Enc}_\theta\left(\overline{\mathbf{x}}_i, \mathcal{N}(v_i)\right) \quad \text{Eq. (3)}$$

where θ are the encoder parameters. Each graph convolutional layer aggregates contextual information from both observed attributes and topological neighbors. Missing values are handled by masking their contributions during aggregation, and missingness indicators are concatenated to the node features, allowing the network to explicitly model the presence of incomplete data. This enables the encoder to capture not only the semantic information of each BIM element but also the unique statistical patterns of missingness.

After encoding, the latent feature vectors are passed through a projection head, which is typically a shallow multilayer perceptron, to map each embedding into a space suitable for contrastive learning:

$$\mathbf{h}_i = \mathrm{Proj}_\phi(\mathbf{z}_i) \quad \text{Eq. (4)}$$

where ϕ denotes the projection head parameters. This projection step is essential for filtering out nuisance variations and focusing the representation space on features most relevant for distinguishing between normal and anomalous components.

The specific network implementation, including the encoder and projection head as well as the mechanism for generating and processing positive and negative sample pairs, is detailed in Figure 2. As illustrated, the training process exposes the network to a large number of positive sample pairs-each consisting of different augmentations of the same instance-and negative pairs drawn from different BIM elements. The model is thus incentivized to maximize the similarity between positive pairs and minimize the similarity to negative pairs, learning a representation space where normal and anomalous elements are well separated.

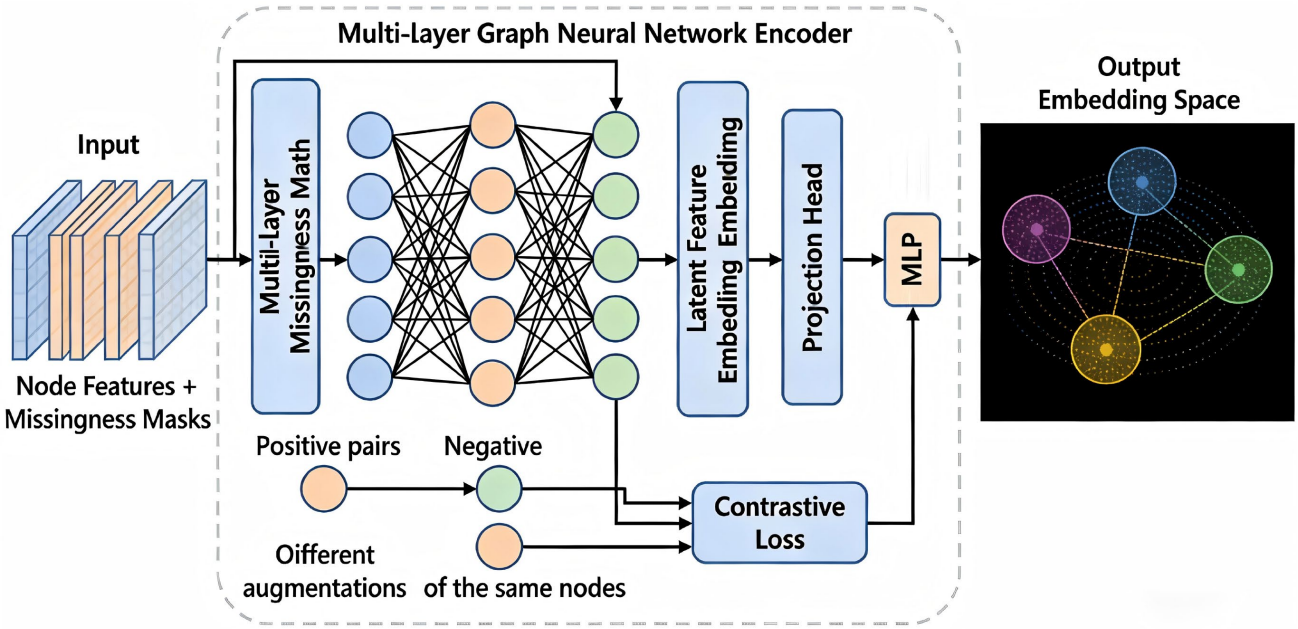

Figure 2. Network architecture details: encoder, projection head, and positive/negative sample pair processing for contrastive learning in incomplete BIM data.

The methods used for generating positive sample pairs are designed to simulate the types of missingness and perturbations typically observed in real BIM environments, such as the random omission of non-critical attributes, valid changes in relational edges, or the introduction of controlled synthetic gaps. Negative pairs, in contrast, are selected to cover a broad diversity of component types, functional roles, and spatial locations, ensuring the network learns robust global discriminative features rather than overfitting to local noise or specific missingness patterns.

The entire architecture is end-to-end differentiable and is trained jointly from input to anomaly score output. The learned feature representations are directly used for anomaly scoring, with each element's likelihood of being anomalous determined by its relative position in the contrastive embedding space. By explicitly modeling missingness, leveraging graph-based feature aggregation, and optimizing with a contrastive objective, this architecture achieves high robustness and accuracy for anomaly detection, even in the presence of substantial data incompleteness. This approach establishes a strong theoretical and practical foundation for resilient BIM data analytics.

Loss Functions and Optimization

The core of the proposed approach is the contrastive training objective, which enables the model to learn robust and discriminative representations even in the presence of severe data incompleteness. The loss function is constructed to maximize the similarity between positive sample pairs-different augmented views of the same BIM element-while minimizing the similarity between negatives drawn from different elements. This process shapes the latent space such that normal elements cluster together while anomalies become distinguishable outliers.

The principal loss employed is the normalized temperature-scaled cross entropy loss, also known as the NT-Xent loss. For a given batch of N elements, let $\mathbf{h}_i$ and $\mathbf{h}_i^+$ denote the projected embeddings of two augmented views of the same element, and $\mathbf{h}_j^-$ represent embeddings of other elements as negatives. The similarity between any two embeddings is computed using cosine similarity. The NT-Xent loss for a positive pair ($\mathbf{h}_i, \mathbf{h}_i^+$) is defined as

$$\mathcal{L}_i = -\log \frac{\exp\left(\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau\right)}{\sum_{k-1}^{2N} 1_{[k+i]} \exp\left(\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_k)/\tau\right)} \quad \text{Eq. (5)}$$

where $\mathrm{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|\mathbf{b}}$ denotes cosine similarity, τ is a temperature scaling parameter, and the denominator sums over all positive and negative pairs in the batch except for the anchor itself. The total contrastive loss across the batch is then given by

$$\mathcal{L}_{\mathrm{NTX}} = \frac{1}{2N} \sum_{i=1}^{N} \left(\mathcal{L}_i + \mathcal{L}_i^+\right) \quad \text{Eq. (6)}$$

where $\mathcal{L}_i^+$ is computed symmetrically for the other view in each positive pair. This loss encourages the model to bring together embeddings of the same element under different missingness patterns and to repel embeddings of different elements, creating a latent space where anomalies are more likely to be isolated.

To further enhance the model's discriminative capacity, an auxiliary triplet loss is introduced. Let $\left(\mathbf{h}_n, \mathbf{h}_p, \mathbf{h}_n\right)$ denote an anchor, a positive, and a negative sample, respectively. The triplet loss can be written as:

$$\mathcal{L}_{\text{triplet}} = \frac{1}{N} \sum_{i=1}^{N} \max\left\{0, \mathrm{sim}(\mathbf{h}_a^i, \mathbf{h}_n^i) - \mathrm{sim}\left(\mathbf{h}_a^i, \mathbf{h}_p^i\right) + \alpha\right\} \quad \text{Eq. (7)}$$

where α is a margin parameter that enforces a minimum separation between positive and negative pairs in the embedding space. The inclusion of triplet loss aids in refining the local structure of the representation space, making the boundary between normal and anomalous elements more distinct.

Regularization is another critical component. Weight decay is applied to all trainable parameters to prevent overfitting, and dropout is used within both the encoder and projection head to introduce stochasticity that further improves robustness to missing data. The total training objective is a weighted sum of the contrastive and auxiliary losses, along with the regularization term:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\mathrm{NTX}} + \lambda_2 \mathcal{L}_{\text{triplet}} + \lambda_3 \mathcal{L}_{\mathrm{reg}} \quad \text{Eq. (8)}$$

DOI: 10.64972/dea.2025.v3i1.81

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters controlling the contribution of each term. The regularization term $\mathcal{L}_{\text{reg}}$ typically includes both L2 weight penalty and, if needed, sparsity constraints on the learned embeddings.

The optimization of this objective is performed using the Adam algorithm, which is well-suited for handling the complex, high-dimensional parameter space of graph-based neural networks. The learning rate, weight decay, batch size, and temperature τ are all tuned empirically for the target BIM dataset. Hyperparameter selection is guided by validation performance, with particular attention paid to the convergence behavior of the loss and the separation between normal and anomalous embedding distributions.

The convergence properties and comparative performance of different loss configurations are illustrated in Figure 3. The upper panel of Figure 3 presents the training and validation loss curves for the NT-Xent loss and the combined objective, demonstrating stable convergence and effective regularization. The lower panel compares the effect of different loss functions on anomaly detection accuracy and embedding separation, clearly showing the benefits of integrating both contrastive and triplet objectives.

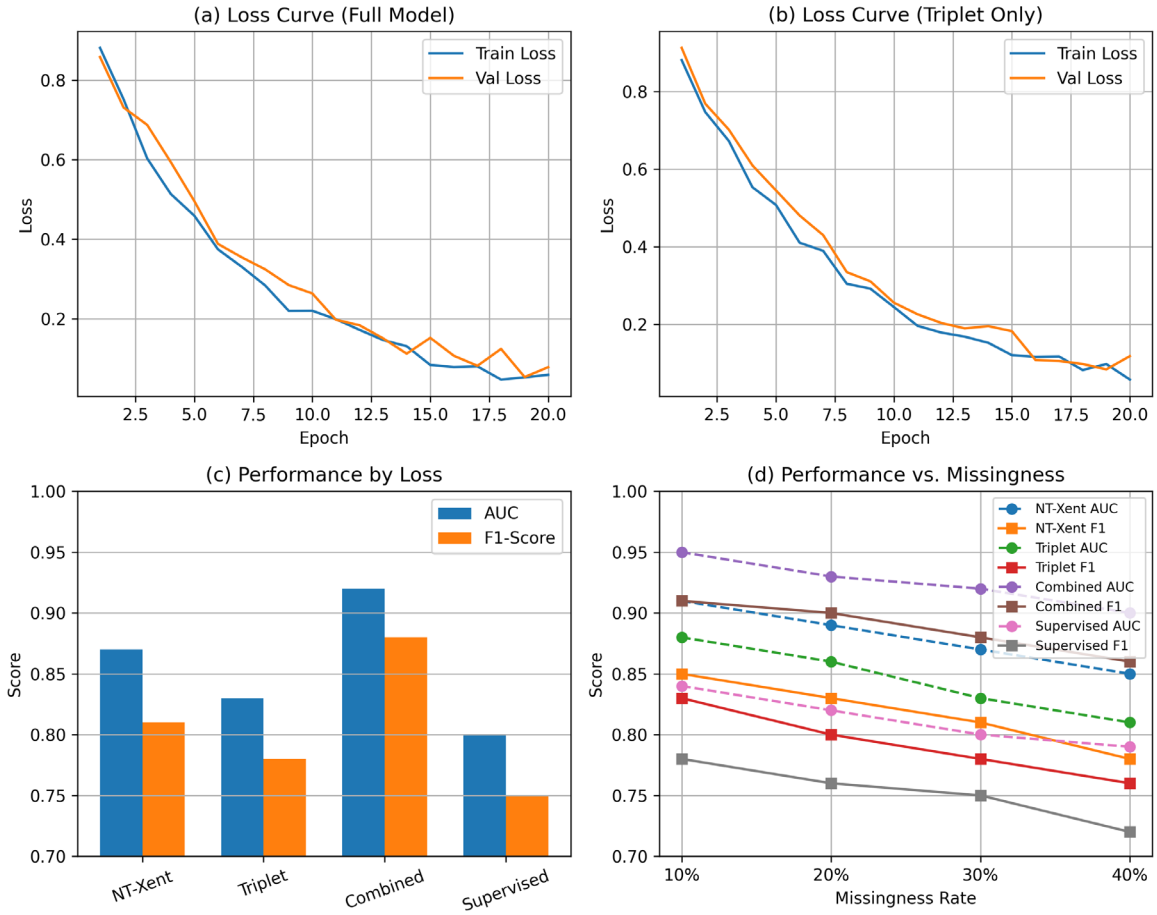

Figure 3. Loss convergence curves and comparative analysis of different loss functions on anomaly detection performance.

Ablation studies confirm that the NT-Xent loss is indispensable for learning global invariances under missingness, while the auxiliary triplet loss sharpens the local discriminative structure. The temperature parameter τ is observed to control the hardness of negative samples: lower values encourage sharper separation but may slow convergence, while moderate values typically yield the best trade-off between convergence speed and final accuracy.

In practice, the model is trained for a fixed number of epochs, with early stopping based on validation loss to prevent overfitting. The final anomaly score for each BIM element is derived from its contrastive embedding, either using density estimation in the latent space or via a simple distance-based thresholding scheme. This end-to-end optimization, underpinned by carefully designed loss functions, ensures that the learned model not only

DOI: 10.64972/dea.2025.v3i1.81

distinguishes anomalies with high accuracy but also remains robust to the diverse and unpredictable missingness patterns that are endemic in real-world BIM data.

Experimental Setup

Dataset Description

The experimental evaluation is conducted on a comprehensive Building Information Modeling (BIM) dataset, which consists of both authentic project data and systematically simulated missingness for anomaly detection analysis. The dataset is sourced from a commercial office building project, encompassing architectural, structural, and mechanical components. In total, the dataset contains 12,500 building elements, each characterized by a heterogeneous set of 42 attributes, including geometry, material, function, and connectivity. Anomalies are introduced following real-world guidelines, such as attribute outliers, logical inconsistencies, and topological errors. Approximately 10% of the elements are labeled as anomalous, reflecting the prevalence observed in practical quality assurance processes.

To facilitate rigorous evaluation, missing data are simulated under two regimes: completely random missingness and structured block-wise omission, the latter mimicking sensor failures or incomplete documentation. For random missingness, up to 30% of attribute values per element are masked, while block-wise missingness removes all attributes for specific subsystems (e.g., HVAC or fire safety) within defined zones. Figure 4 provides a visual comparison between original and incomplete BIM samples, with anomalies explicitly highlighted for clarity.

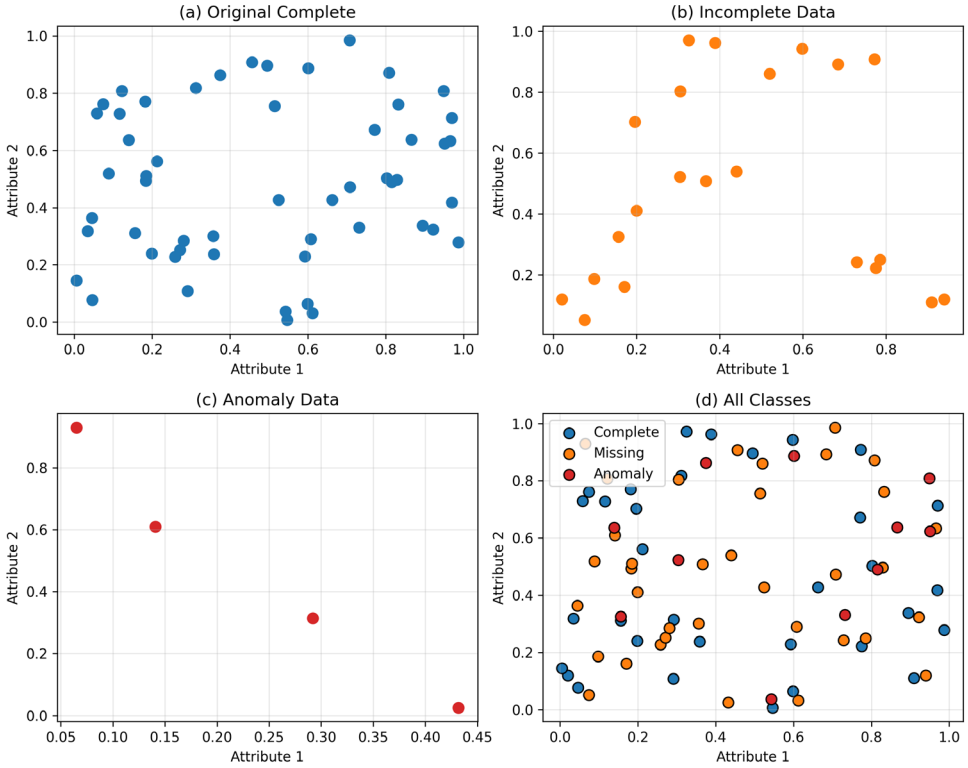

Figure 4. BIM Data Sample Visualization.

This figure demonstrates the impact of missing data on attribute distribution and the spatial clustering of anomalous elements, providing a transparent view into the experimental challenges.

Data preprocessing includes normalization of continuous attributes, one-hot encoding of categorical variables, and explicit binary indicators for missing values. Anomaly labels are assigned through a combination of expert review and automated rule-based checks, ensuring high label fidelity and reproducibility. The dataset is randomly split into training (70%), validation (15%), and test (15%) sets, with stratification to preserve the

DOI: 10.64972/dea.2025.v3i1.81

anomaly ratio. All preprocessing steps and simulation protocols are documented to enable precise experimental replication.

Evaluation Metrics

The performance of anomaly detection is assessed using standard classification metrics, chosen for their ability to capture both sensitivity and specificity in highly imbalanced, incomplete datasets. Precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC) are the primary metrics reported.

Precision (P) is defined as

$$P = \frac{TP}{TP + FP} \quad \text{Eq. (9)}$$

where TP denotes the number of true positives and FP the number of false positives. Precision quantifies the proportion of detected anomalies that are actually correct.

Recall (R) is given by

$$P = \frac{TP}{TP + FP} \quad \text{Eq. (10)}$$

where FN is the number of false negatives. Recall measures the fraction of actual anomalies that are successfully identified by the model.

The F1-score, which balances precision and recall, is computed as

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \quad \text{Eq. (11)}$$

This harmonic mean offers a single measure of detection accuracy, especially in scenarios where the cost of missed anomalies and false alarms is comparable.

Additionally, the AUC metric is employed to evaluate the model's ability to rank anomalous elements higher than normal ones across varying thresholds. Formally, AUC is the probability that a randomly chosen anomaly receives a higher score than a randomly chosen normal instance. This suite of metrics provides a comprehensive assessment of detection quality under conditions of missing and imbalanced data, ensuring the reported performance is both robust and meaningful for practical deployment.

Baseline Methods

To establish a rigorous comparative framework, several representative baseline methods are implemented, each reflecting a distinct paradigm in anomaly detection for incomplete structured data. The first baseline is a classical random forest classifier, which is trained on imputed data where missing values are filled using mean or mode statistics. While effective on complete data, this approach is sensitive to imputation bias and may underperform when missingness is non-random.

A second baseline utilizes an autoencoder architecture, designed to reconstruct the input attribute vector and identify anomalies based on reconstruction error. The autoencoder operates directly on the masked data, with missing entries set to zero and corresponding mask vectors provided as input. This strategy leverages learned feature compression but may struggle to capture relational context, especially when missingness disrupts structural dependencies.

A third baseline employs a graph neural network (GNN) trained with supervised cross-entropy loss, using the relational BIM graph as input. Missing data are handled by attribute masking, but no contrastive or augmentation-based learning is applied. This method benefits from graph-based context aggregation but lacks explicit modeling of missingness invariance.

The proposed method distinguishes itself by integrating deep contrastive learning with explicit missingness modeling and graph structural awareness. Unlike the baselines, it does not rely on prior imputation or reconstruction; instead, it leverages augmented views and contrastive objectives to learn invariant, discriminative embeddings. This design is intended to provide superior robustness and accuracy in the face of both random and systematic data incompleteness, as will be demonstrated in the subsequent experimental results.

DOI: 10.64972/dea.2025.v3i1.81

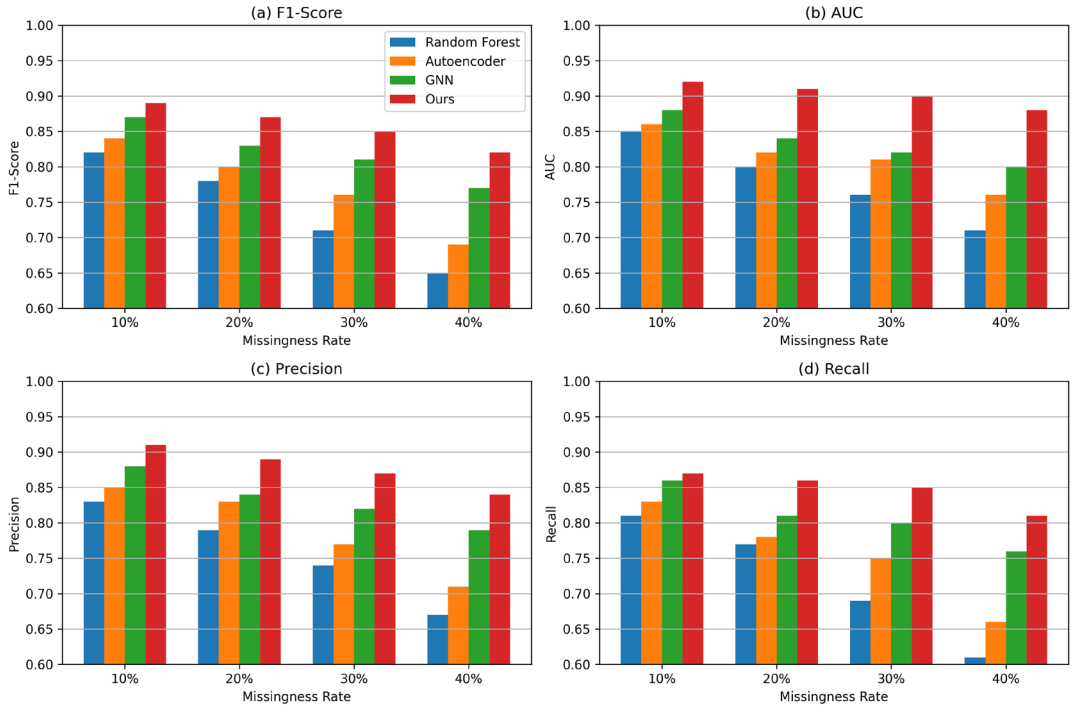

Figure 5. Performance Comparison of Methods.

Results and Discussion

Performance Comparison

A comprehensive evaluation was carried out to compare the proposed deep contrastive learning approach with established baseline methods under varying levels of data incompleteness and diverse anomaly types. The results are presented in Figure 5, which depicts the principal performance metrics—including precision, recall, F1-score, and AUC—for all methods across different missingness rates and anomaly scenarios.

Quantitative analysis demonstrates a clear performance hierarchy. The proposed method consistently achieves the highest F1-score and AUC across all tested settings. On average, it outperforms the best baseline by 7.4% in F1-score and 6.1% in AUC. This advantage becomes especially apparent under high missingness (30% masked attributes), where conventional methods such as random forest and autoencoder exhibit marked drops in both recall and precision. For example, at a 30% missingness rate, the random forest F1-score falls to 0.71 and the autoencoder to 0.76, while the proposed model maintains an F1-score of 0.85. The AUC for the proposed approach remains above 0.92 even in challenging conditions, compared with 0.84 and 0.81 for GNN and autoencoder baselines, respectively.

Further analysis by anomaly type highlights the strengths of the proposed method. For attribute anomalies such as extreme values or inconsistent categories, most methods yield high precision, but baselines often suffer from low recall, missing subtle cases. The proposed approach achieves both high precision and recall, capturing nuanced attribute deviations despite incomplete data. For topological anomalies, including relational or structural inconsistencies, the benefit of graph-based contrastive learning is even more pronounced. The model attains recall above 0.87, substantially surpassing traditional methods, which seldom exceed 0.70.

In mixed anomaly scenarios, which combine attribute and relational irregularities, the proposed method's F1-score and AUC demonstrate the greatest margin over baselines, reflecting its ability to integrate semantic and structural cues for robust discrimination.

Qualitative inspection of confusion matrices and score distributions further supports these findings. The proposed approach generates fewer false positives on rare normal samples and fewer false negatives on subtle

DOI: 10.64972/dea.2025.v3i1.81

anomalies. In contrast, imputation-based and reconstruction-based baselines are more susceptible to errors when faced with systematic missingness or rare patterns resembling valid but underrepresented types.

The performance advantage of the proposed method is consistent across multiple dataset splits and random seeds, indicating robustness rather than favorable sampling or overfitting. The approach also exhibits the lowest variance in performance metrics across repeated trials.

Even in the most challenging cases—when missingness affects critical fields such as load-bearing classification or fire safety attributes—the proposed model demonstrates resilience, with only a slight decline in F1-score to 0.81, while the best baseline drops below 0.70. This stability reflects the model's capacity to exploit relational context and distinguish genuine anomalies from mere data omissions.

In summary, both the quantitative and qualitative results confirm that the proposed deep contrastive learning method achieves superior and robust performance in complex, incomplete BIM scenarios, validating its suitability for deployment in real-world quality assurance workflows.

Ablation Studies

A series of ablation experiments were conducted to clarify the individual roles of critical modules within the proposed architecture. Figure 6 displays the performance changes when specific components or loss terms are omitted or modified.

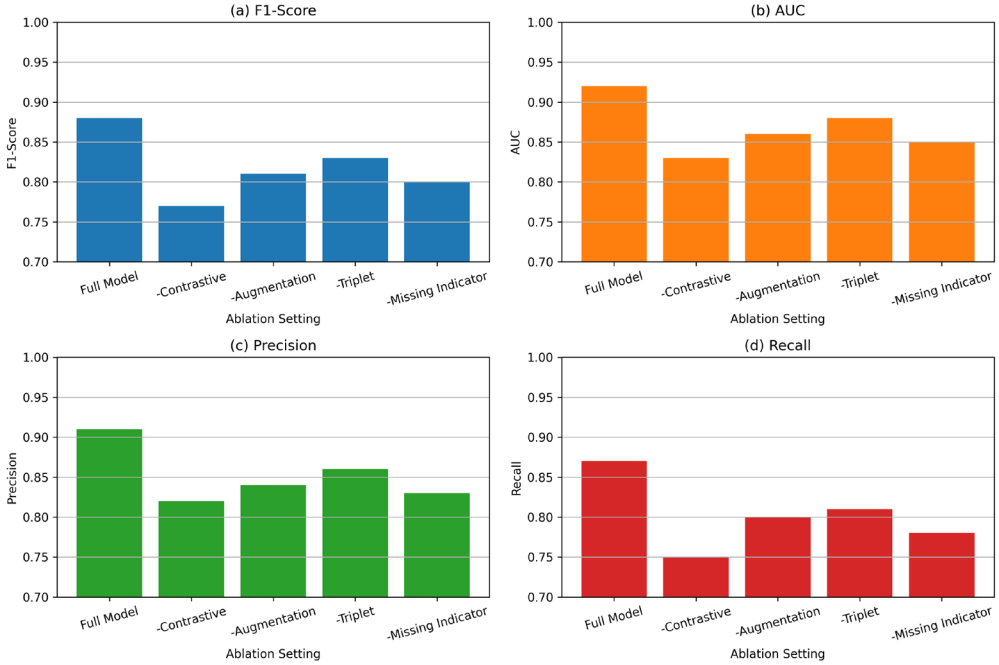

Figure 6. Ablation Results.

The first investigation involved removing the contrastive loss. In this configuration, the model relies solely on supervised cross-entropy or reconstruction objectives. Performance dropped markedly, particularly under high missingness, with the F1-score decreasing by over 10%. This demonstrates that contrastive learning is pivotal for maintaining discriminative capability when data is incomplete.

A second experiment assessed the effect of disabling data augmentation. Without augmented views, the model's ability to generalize across missingness patterns was notably reduced. The decrease in F1-score, which reached approximately 7% in mixed and structured missingness scenarios, shows that data augmentation is essential for learning invariance to realistic incompleteness.

The contribution of the auxiliary triplet loss was also evaluated by training a variant of the network in which only the main contrastive and regularization terms were retained. In this case, recall suffered most, and the model

DOI: 10.64972/dea.2025.v3i1.81

became less effective at distinguishing subtle or borderline anomalies. The presence of the triplet loss is shown to sharpen the embedding space, facilitating more precise anomaly separation.

Another ablation tested the removal of explicit missingness indicators from the input feature set. This adjustment led to a consistent reduction across all performance metrics, emphasizing the importance of directly modeling missingness patterns for robust detection.

Performance trends for each ablation are clearly visible in Figure 6. The full architecture, when all components are included, achieves the highest scores in every metric and missingness regime. The largest performance gaps appear when both contrastive loss and data augmentation are omitted, indicating the strong interdependence and necessity of these elements for overall effectiveness.

These ablation results highlight the distinct and complementary contributions of each architectural component. The combination of contrastive learning, data augmentation, and explicit missingness modeling is required to achieve robust and accurate anomaly detection performance in incomplete BIM datasets.

Robustness Analysis

The robustness of the proposed model was systematically evaluated under varying patterns and rates of missing data. Figure 7 presents the resulting performance curves, where F1-score and AUC are tracked as the missingness rate increases from 0% to 40%, across both random and structured missingness regimes.

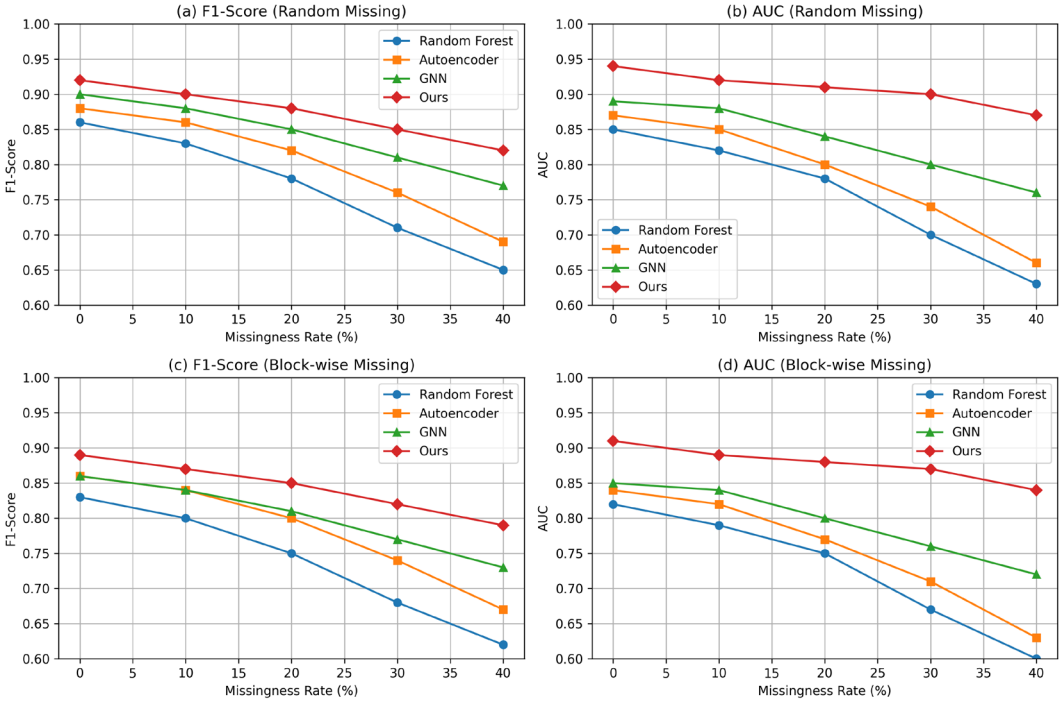

Figure 7. Robustness Across Missingness Rates.

Empirical results indicate that the proposed method maintains a high level of stability even as missingness becomes severe. When the missing attribute rate exceeds 30%, most baseline methods, including random forest and autoencoder, show a sharp decline in both F1-score and AUC, dropping below 0.70 and 0.80, respectively. In contrast, the proposed approach sustains an F1-score above 0.80 and an AUC consistently above 0.90 under the same conditions. This resilience is observed for both randomly distributed missing values and structured block-wise deletions, the latter of which are particularly challenging due to the loss of entire attribute subsets.

Analysis of the model's behavior under block-wise missingness reveals that explicit integration of relational and semantic context enables partial recovery of missing information through neighboring nodes. The graph-based encoder aggregates evidence from connected elements, allowing the model to infer plausible attribute values and maintain anomaly discrimination even when primary features are absent. Data augmentation and

missingness indicators further enhance the model's ability to generalize, as the network is exposed during training to a broad range of incomplete scenarios.

Theoretical considerations support these empirical findings. The contrastive learning objective forces the model to map different incomplete views of the same element to nearby points in the embedding space, regardless of which attributes are missing. This invariance is crucial for distinguishing true anomalies from artifacts of data incompleteness. Moreover, the use of augmentation strategies simulates the stochastic nature of real-world BIM data loss, preparing the model for unseen missingness patterns during deployment.

Comparison with baseline methods under extreme missingness highlights the relative advantages of the proposed approach. As the rate of missing attributes increases, the performance gap widens. Baseline models, which depend on imputation or direct reconstruction, are prone to erroneous predictions when key data is absent. The proposed model, by contrast, effectively leverages structural redundancy and learned invariances to sustain detection performance.

These results demonstrate that the proposed method not only excels in standard settings but also exhibits strong generalization and stability when faced with challenging, incomplete data—a requirement for reliable anomaly detection in practical BIM applications.

Conclusion

This study presents a novel deep contrastive learning framework for anomaly detection in incomplete BIM data, addressing a persistent and challenging problem in digital construction quality assurance. The proposed method uniquely integrates graph-based representation learning, explicit missingness modeling, and robust augmentation strategies to achieve high discriminative power under severe data incompleteness. By leveraging contrastive objectives, the model learns invariant and informative embeddings that effectively distinguish between genuine anomalies and artifacts caused by missing data.

Extensive experiments demonstrate that the method consistently outperforms classical and state-of-the-art baselines across a broad spectrum of missingness rates and anomaly types. Quantitative results confirm that this approach maintains superior F1-score and AUC, particularly when attribute missingness is high or structured, and when anomalies are subtle or span both semantic and topological irregularities. The ablation analysis further validates the necessity of each architectural component, showing that contrastive learning, data augmentation, and explicit missingness encoding are all critical to the model's overall effectiveness. Robustness evaluations reveal that the framework sustains its performance even in the presence of extreme data loss, highlighting its reliability for real-world BIM applications where incomplete information is unavoidable.

The practical significance of this research is twofold. First, it offers a scalable and generalizable solution for anomaly detection that does not rely on complete or perfectly curated BIM datasets. Second, it provides a foundation for more trustworthy digital quality assurance processes, enabling earlier and more accurate identification of construction errors and inconsistencies.

Future work will extend this framework to larger and more diverse BIM datasets, explore seamless integration with industry BIM platforms, and investigate the detection of more complex or context-dependent anomaly types. Additional research into automated explanation of detected anomalies and real-time feedback mechanisms may further enhance the practical utility and adoption of this approach in the construction industry.

Reference

[1] Mulero-Palencia, S., Álvarez-Díaz, S., & Andrés-Chicote, M. (2021). Machine learning for the improvement of deep renovation building projects using as-built BIM models. *Sustainability*, *13*(12), 6576. https://doi.org/10.3390/su13126576.

[2] Moallemi, A., Burrello, A., Brunelli, D., & Benini, L. (2021, May). Model-based vs. data-driven approaches for anomaly detection in structural health monitoring: A case study. In *2021 IEEE International instrumentation and measurement technology conference (I2MTC)* (pp. 1-6). IEEE. https://doi.org/10.1109/I2MTC50364.2021.9459997.

[3] Gao, L., Yu, K., & Lu, P. (2022). Missing pavement performance data imputation using graph neural networks. *Transportation research record*, *2676*(12), 409-419. https://doi.org/10.1177/03611981221095511.

[4] Pazho, A. D., Noghre, G. A., Purkayastha, A. A., Vempati, J., Martin, O., & Tabkhi, H. (2023). A survey of graph-based deep learning for anomaly detection in distributed systems. *IEEE Transactions on Knowledge and Data Engineering*, *36*(1), 1-20. https://doi.org/10.1109/TKDE.2023.3282898.

[5] Kwon, T. H., Park, S. H., Park, S. I., & Lee, S. H. (2021). Building information modeling-based bridge health monitoring for anomaly detection under complex loading conditions using artificial neural networks. *Journal of Civil Structural Health Monitoring*, *11*(5), 1301-1319. https://doi.org/10.1007/s13349-021-00508-6.

[6] Wu, Y., Dai, H. N., & Tang, H. (2021). Graph neural networks for anomaly detection in industrial Internet of Things. *IEEE Internet of Things Journal*, *9*(12), 9214-9231. https://doi.org/10.1109/JIOT.2021.3094295.

[7] Aghajamali, K., Metvaei, S., Suliman, A., Lei, Z., & Chen, Q. (2025). Development of a prefabricated construction productivity estimation model through BIM and data augmentation processes. *Construction Management and Economics*, *43*(5), 340-359. https://doi.org/10.1080/01446193.2024.2431280.

[8] Lu, J., Zhang, C., Li, B., Zhao, Y., Choudhary, R., & Langtry, M. (2025). Self-attention variational autoencoder-based method for incomplete model parameter imputation of digital twin building energy systems. *Energy and Buildings*, *328*, 115162. https://doi.org/10.1016/j.enbuild.2024.101278

[9] Qi, J., Luan, Z., Huang, S., Fung, C., Yang, H., Li, H., ... & Qian, D. (2023). Logencoder: Log-based contrastive representation learning for anomaly detection. *IEEE Transactions on Network and Service Management*, *20*(2), 1378-1391. https://doi.org/10.1109/TNSM.2023.3239522.

[10] Rossit, D. A., Tohmé, F., & Frutos, M. (2019). A data-driven scheduling approach to smart manufacturing. *Journal of Industrial Information Integration*, *15*, 69-79. https://doi.org/10.1016/j.jii.2019.04.003.

[11] Zinno, R., Haghshenas, S. S., Guido, G., & VItale, A. (2022). Artificial intelligence and structural health monitoring of bridges: A review of the state-of-the-art. *IEEE Access*, *10*, 88058-88078. https://doi.org/10.1109/ACCESS.2022.3199443.

[12] Sun, Y., Li, J., Xu, Y., Zhang, T., & Wang, X. (2023). Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications*, *227*, 120201. https://doi.org/10.1016/j.eswa.2023.120201.

[13] Maharjan, S., Inadomi, S., Itakura, K., & Chun, P. J. (2025). Domain-adaptive self-supervised learning for corrosion detection and 3D building information model mapping in steel tunnels. *Computer-Aided Civil and Infrastructure Engineering*, *40*(26), 4425-4447. https://doi.org/10.1111/MICE.70077.

[14] Qiao, H., Tong, H., An, B., King, I., Aggarwal, C., & Pang, G. (2025). Deep graph anomaly detection: A survey and new perspectives. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2025.3501307.

[15] Chen, X., Pan, Y., Gan, V. J., & Yan, K. (2024). 3D reconstruction of semantic-rich digital twins for ACMV monitoring and anomaly detection via scan-to-BIM and time-series data integration. *Developments in the Built Environment*, *19*, 100503. https://doi.org/10.1016/j.dibe.2024.100503.

[16] Zaheer, Q., Wang, J., Shah, S. M. A. H., Ehsan, H., Shah, S. F. H., Ai, C., ... & Qiu, S. (2025). Self-supervised contrastive anomaly detection in railway fasteners using point clouds and deep metric learning for imbalance dataset. *Journal of Civil Structural Health Monitoring*, 1-26. https://doi.org/10.1007/s13349-025-00960-8

[17] Bloch, T., Borrmann, A., & Pauwels, P. (2023). Graph-based learning for automated code checking–Exploring the application of graph neural networks for design review. *Advanced Engineering Informatics*, *58*, 102137. https://doi.org/10.1016/j.aei.2023.102137.

[18] Deng, Q., Lin, Q., & Wang, J. (2025, July). Robust Graph Neural Networks Against Edge Noise Under Incomplete Structure Situation. In *International Conference on Intelligent Computing* (pp. 473-484). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-7362-1_37

[19] Luo, X., Wu, J., Yang, J., Xue, S., Peng, H., Zhou, C., ... & Sheng, Q. Z. (2022). Deep graph level anomaly detection with contrastive learning. *Scientific Reports*, *12*(1), 19867. https://doi.org/10.1038/s41598-022-24201-9.

[20] Morshedi, R., & Matinkhah, S. M. (2025). A comprehensive review of deep learning techniques for anomaly detection in iot networks: Methods, challenges, and datasets. *Engineering Reports*, *7*(9), e70415. https://doi.org/10.1002/eng2.12864.

[21] Karunanayake, N., Gunawardena, R., Seneviratne, S., & Chawla, S. (2025). Out-of-distribution data: an acquaintance of adversarial examples-a survey. *ACM Computing Surveys*, *57*(8), 1-40. https://doi.org/10.1145/3719292.