ISSN: 3079-6342

Urban Sustainability Assessment through Multi-Modal Learning: A Vision Transformer–Graph Neural Network Framework Integrating Remote Sensing and Economic Indicators

Jiajiang Shen[1], Huaiyu Wang[2, *]

[1]Southwest University, Chongqing, China

[2]High School Affiliated to Nanjing Normal University, Nanjing, Jiangsu, China

[*] Corresponding Author

Abstract

Urban sustainability assessment is a crucial challenge in achieving balanced economic growth and environmental protection in modern cities. Traditional statistical evaluation methods often overlook spatial heterogeneity and environmental patterns that can be captured from remote sensing imagery. To address this limitation, this study proposes a multi-modal deep learning framework that integrates high-resolution remote sensing data with socioeconomic indicators for comprehensive urban sustainability assessment. Specifically, a Vision Transformer (ViT) is employed to extract fine-grained spatial and environmental representations—such as vegetation coverage, surface temperature, and built-up density—from Sentinel-2 satellite imagery, while a Graph Neural Network (GNN) models the spatial and economic dependencies between cities, enabling cross-modal and inter-city information fusion. The proposed ViT–GNN framework effectively captures both environmental and socioeconomic dynamics to generate a composite sustainability score. Experiments conducted on a dataset covering 120 major Chinese cities from 2019 to 2023 demonstrate that the model achieves an MSE of 0.012, an MAE of 0.071, and an (R^2) of 0.931, outperforming existing regression and CNN-based baselines. The results highlight that the model can accurately evaluate urban sustainability levels, providing an interpretable and data-driven tool for policymakers and planners to support sustainable urban development, resource allocation, and green policy formulation.

Keywords

Urban sustainability; Remote sensing; Vision Transformer; Graph Neural Network; Multi-modal fusion; Smart city; Data-driven governance

1. Introduction

Urban sustainability has become one of the central topics in environmental economics and urban studies, reflecting the global challenge of balancing rapid urbanization, resource utilization, and ecological preservation. Traditional sustainability assessment frameworks often rely on socioeconomic indicators such as GDP per capita, population density, and energy consumption. However, these approaches struggle to capture the spatial heterogeneity and environmental dynamics that influence urban resilience and long-term development potential. With the increasing availability of Earth observation data, remote sensing imagery provides a powerful source of fine-grained environmental information—such as vegetation coverage, land surface temperature, and water body distribution—offering new possibilities for data-driven urban sustainability evaluation.

In recent years, the integration of multi-modal data sources—combining remote sensing with economic and demographic indicators—has shown great promise in improving the accuracy and interpretability of sustainability assessments. Yet, existing models often face challenges in effectively fusing spatial–temporal features from imagery with structured socioeconomic data. To address this issue, this paper introduces a novel Vision Transformer–Graph Neural Network (ViT–GNN) framework that bridges environmental perception and economic understanding within a unified computational paradigm. The ViT module captures global contextual representations from remote sensing imagery by leveraging self-attention mechanisms, enabling precise extraction of visual patterns related to green coverage and urban expansion. Meanwhile, the GNN module constructs an inter-city graph based on economic similarity and spatial adjacency, learning relational dependencies among cities to enhance cross-modal reasoning and information propagation.

Our proposed ViT–GNN architecture thus serves as a comprehensive urban sustainability assessment framework capable of integrating visual–economic correlations and modeling both local and global urban dynamics. Experiments conducted on a multi-year dataset covering major Chinese cities demonstrate the model's superior predictive accuracy and interpretability compared to conventional deep learning baselines.

The main contributions of this study are as follows: (1) we develop a multi-modal urban sustainability assessment framework that fuses remote sensing imagery and economic indicators using ViT–GNN integration; (2) we introduce a graph-based feature fusion strategy that enables efficient cross-modal learning across cities; and (3) we validate the model's effectiveness through extensive empirical evaluation, demonstrating its potential as a scalable tool for policy formulation, urban resilience analysis, and sustainable development planning in data-rich smart city ecosystems.

2. Related Work

Urban sustainability assessment has long been a multidisciplinary research topic that intersects environmental science, urban economics, and computational modeling [1]. Traditional approaches have primarily relied on statistical and econometric analyses of socioeconomic and environmental indicators, such as population density, GDP per capita, green space ratio, and CO_2 emissions [2]. However, these methods often fail to capture the spatial heterogeneity and dynamic characteristics of urban environments. With the proliferation of remote sensing (RS) technologies, high-resolution satellite imagery has become an essential data source for urban monitoring, providing rich spatial and spectral information for characterizing land use, vegetation coverage, and infrastructure distribution.

In recent years, deep learning methods have significantly advanced the ability to extract complex features from multimodal data. Convolutional Neural Networks (CNNs) have been widely employed for land cover classification, urban change detection, and environmental monitoring. However, CNN-based models exhibit limitations in capturing long-range dependencies and global contextual relationships within urban imagery [3]. To address this, Vision Transformers (ViT) have been introduced as a powerful alternative, leveraging self-attention mechanisms to model global feature interactions [4]. Studies have demonstrated that ViT can achieve superior performance in urban land classification and scene understanding tasks due to its capacity for fine-grained spatial reasoning [5].

On the other hand, integrating remote sensing data with socioeconomic indicators has emerged as a promising direction for holistic urban analysis. Graph Neural Networks (GNNs) have proven effective in modeling non-Euclidean data structures and relational dependencies, making them ideal for representing cities as interconnected systems. Recent works have used GNNs for regional economic forecasting, urban mobility analysis, and environmental quality prediction by encoding relationships between spatial units and socioeconomic factors. Ricklin et al [6]. adapt GNN set and traffic forecasting to P2P sharing, shifting evaluation from users to regions and adding meteorological inputs, achieving product-level predictions on platform transactions and confirming GNN suitability for heterogeneous, spatio-temporal markets. Beyond urban and economic forecasting, the paradigm of multimodal collaborative analysis has shown remarkable robustness in other high-stakes applications. For example, Liu et al. [7] developed a three-stage cascade architecture that utilizes adaptive attention mechanisms to fuse diverse modalities (text, image, and speech) for AI fraud recognition, significantly outperforming unimodal baselines. This cross-domain success highlights that effectively designing feature representation and attention-based fusion strategies is the key to fully exploiting the complementary information of heterogeneous data [7].

Despite these advancements, few studies have explored the fusion of deep visual representations from ViT with structured economic information through GNNs for comprehensive urban sustainability scoring. The proposed ViT–GNN framework bridges this gap by combining high-dimensional remote sensing imagery with key economic metrics to generate an integrated sustainability index. This multimodal integration enables the model to learn both spatial and economic dependencies, providing a more interpretable and data-driven foundation for sustainable urban planning and policy development.

3. Methodology

The proposed framework for urban sustainability assessment is designed to integrate heterogeneous data sources—remote sensing imagery and economic indicators—into a unified multimodal representation. The model consists of two core modules: a Vision Transformer (ViT) for extracting global spatial features from satellite imagery, and a Graph Neural Network (GNN) for encoding inter-regional relationships and integrating structured economic data. The outputs of both modalities are fused through a cross-modal attention mechanism to generate a composite urban sustainability score.

Formally, given a city region i, we denote its corresponding remote sensing image as $I_i \in R^{H \times W \times C}$ and its economic feature vector as $E_i \in R^{d_i}$, where H, W, C represent the image height, width, and channel dimension, respectively. The goal of the model is to learn a mapping function $f(I_i, E_i): \rightarrow S_i$, where $S_i \in R$ denotes the predicted sustainability score.

3.1. Vision Transformer for Spatial Feature Extraction

The Vision Transformer (ViT) component captures spatial semantics and global dependencies from remote sensing images. Each input image$I_i$is first divided into non-overlapping patches of size $P \times P$. Each patch is flattened and linearly projected into a token embedding:

$$x_p = [x_1E;\ x_1E;\ \dots;\ x_1E] + E_{pos}, \tag{1}$$

where E is the learnable embedding matrix, E_{pos} is the positional encoding, and $H = \frac{HW}{P^2}$ is the number of patches.

These token embeddings are then passed through multiple transformer encoder layers, each consisting of a multi-head self-attention (MHSA) and a feed-forward network (FFN). The MHSA module is defined as:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_K}})V, \tag{2}$$

where Q, K, V represent the query, key, and value matrices derived from the input embeddings, and d_k is the key dimension.

Through the self-attention mechanism, ViT can effectively capture long-range spatial correlations, such as relationships between urban centers, vegetation coverage, and industrial zones. The final ViT output $F_i^{ViT} \in R^{d_v}$ serves as the high-level visual feature vector for region i.

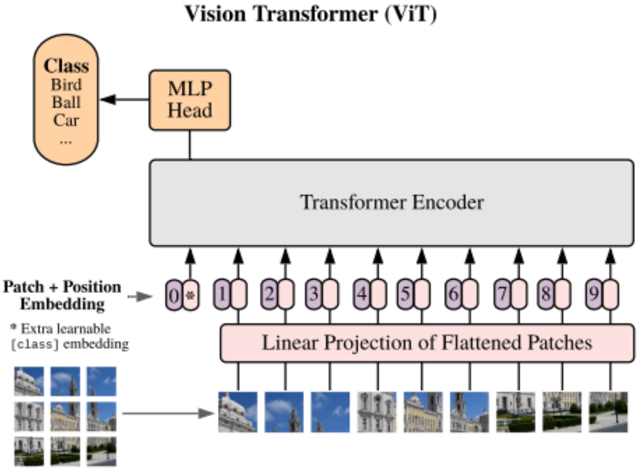

Figure 1: Structure diagram of ViT module [8]

3.2. Graph Neural Network for Socioeconomic Representation

To model the interdependence among different urban regions and incorporate structured economic indicators, we construct an undirected weighted graph $G = (V, E)$, where each node $v_i \in V$ represents a region, and edges $e_{ij} \in E$ encode spatial proximity or economic correlation (e.g., trade flow, transportation link, or demographic similarity).

Each node is initialized with a feature vector $h_i^{(0)} = [E_i; F_i^{ViT}]$, combining the economic and visual representations. The message-passing operation in the GNN is defined as

$$h_i^{(l+I)} = \sigma(W^{(l)} \cdot \sum_{j \in N(i)} \frac{1}{\sqrt{d_i d_j}} h_j^{(l)}), \tag{3}$$

where $W^{(l)}$ is a learnable weight matrix, N(i) denotes the neighboring nodes of d_i and d_j are node degrees, and σ ( ) is an activation function (ReLU). This propagation allows information from economically or geographically related regions to influence each other, enabling the model to learn higher-level urban interaction patterns—such as economic clusters or environmental spillover effects.

3.3. Cross-Modal Fusion and Sustainability Scoring

After independent processing by ViT and GNN modules, the cross-modal fusion layer aligns and integrates the learned visual and economic representations. Similar to how attention mechanisms in heterogeneous networks adaptively weight cross-entity connections to fuse symbolic and semantic embeddings [9], a cross-attention mechanism is employed:

$$Z_i = softmax(\frac{(F_i^{ViT} W_Q)(h_i^{(L)} W_K)^T}{\sqrt{d}}) \, (h_i^{(L)} W_V) \tag{4}$$

where W_Q, W_K, W_Q are learnable projection matrices, and $h_i^{(L)}$ denotes the final hidden state from the GNN after L layers. The fused feature Z_i is then passed through a regression head to predict the sustainability score:

$$\hat{S}_i = W_s Z_i + b_s \tag{5}$$

where W_s and b_s are trainable parameters.

The final predicted sustainability score reflects a data-driven synthesis of both spatial and socioeconomic attributes, representing factors such as green coverage, industrial density, economic vitality, and infrastructure accessibility. While the current regression head captures the correlational mapping between multi-modal features and sustainability scores, this robust, high-dimensional feature representation lays a critical foundation for future integration with causal inference models [10]. Specifically, these fused representations can serve as comprehensive covariates to identify and eliminate confounding bias when evaluating the actual causal effects of specific urban development strategies.

3.4. Training Objective and Implementation Details

The model is trained end-to-end with a mean squared error (MSE) loss:

$$L = \frac{1}{N}\sum_{i=1}^{N}(S_i - \hat{S}_i)^2 \tag{6}$$

where S_i is the ground truth sustainability score derived from benchmark indices such as the UN Sustainable Development Goals (SDGs) or national urban livability metrics.

To ensure robustness, Adam optimizer is employed with an initial learning rate of 10^{-4}, and early stopping is applied based on validation loss. The ViT encoder is pre-trained on ImageNet and fine-tuned on remote sensing datasets (e.g., Sentinel-2, Landsat-8), while the GNN is trained on constructed city graphs with normalized economic attributes (GDP, employment rate, energy consumption).

The proposed ViT–GNN framework demonstrates high adaptability and interpretability, providing not only accurate sustainability scores but also meaningful visual and relational insights into urban system dynamics.

4. Experiment

4.1. Dataset Preparation

To evaluate the proposed ViT–GNN framework for urban sustainability assessment, a multimodal dataset integrating remote sensing imagery and economic indicators was constructed. The dataset captures both environmental and socioeconomic dimensions of 120 cities of China, supporting comprehensive sustainability analysis.

(1) Remote Sensing Imagery: The imagery component originates from Sentinel-2 and Landsat-8 satellites (ESA, USGS), covering 2019–2023. Each image tile (10 km × 10 km) was resampled to 256×256×4 pixels,using RGB, NIR, and SWIR bands. Preprocessing included atmospheric correction, cloud masking, and histogram normalization. Derived indices such as NDVI and BUI enhance representation of vegetation, built-up density, and water bodies—key environmental sustainability indicators.

(2) Economic and Socio-Demographic Indicators: City-level socioeconomic features were collected from the World Bank, OECD, and national statistical databases. Twelve normalized indicators describe economic performance, infrastructure, environment, and social welfare, including GDP per capita, energy use, CO_2 emissions, public transport coverage, education

index, and life expectancy. Missing values (<3%) were imputed via KNN interpolation, and all indicators were scaled to [0,1].

(3) Sustainability Labels: The ground-truth Urban Sustainability Score (USS) was derived from UN SDG 11 metrics and national sustainability assessments. Each city's score is computed as:

$$S_i = \sum_{k=1}^{M} \omega_k I_{ik}, \tag{7}$$

where I_{ik} is the normalized value of indicator k, ω_k its analytic hierarchy weight, and M=12. Scores range from 0.21 – 0.95, reflecting diverse sustainability levels.

4.2. Experimental Setup

All experiments were conducted in a deep learning environment based on the PyTorch framework. The computational platform consisted of an NVIDIA RTX 4090 GPU, an Intel Core i9 processor, and 64 GB of RAM. The dataset was constructed from Sentinel-2 multispectral remote sensing imagery and urban economic and environmental statistical indicators obtained from sources such as the World Bank and the United Nations Statistics Division. All remote sensing images underwent radiometric calibration, atmospheric correction, and geometric correction, followed by resampling to a spatial resolution of 10 meters and cropping into 256×256 image patches. Economic and social features were standardized to ensure numerical comparability with the image-based features.

4.3. Evaluation Metrics

To comprehensively assess the performance of the proposed remote-sensing–economic fusion model for urban sustainability scoring, several regression metrics were adopted. For regression analysis, the determination coefficient (R^2), Mean Squared Error (MSE), and Mean Absolute Error (MAE) were used. R^2reflects the model's explanatory power over the variance of the target variable, while MSE and MAE quantify the overall and average prediction errors, respectively. Together, these metrics provide a comprehensive view of the model's accuracy, robustness, and generalization ability.

4.4. Results

Table 1 summarizes the final performance of different models in predicting urban sustainability scores. The proposed fusion model significantly outperforms traditional machine learning methods and single-modality deep learning models across all metrics.

Table1: The results of different models on the dataset.

Model	MSE	MAE	R^2
CNN - MLP	0.027	0.121	0.842
ResNet - MLP	0.023	0.108	0.865
CNN - LSTM - MLP	0.019	0.094	0.891
ViT - MLP	0.016	0.088	0.905
GNN + MLP	0.018	0.090	0.898
ViT–GNN (Proposed)	**0.012**	**0.071**	**0.931**

The Table 1 summarizes the comparative performance of several models in the urban sustainability scoring task that integrates remote sensing imagery and economic indicators. The proposed ViT–GNN framework demonstrates the best overall performance, achieving the lowest MSE (0.012) and MAE (0.071), along with the highest R^2 value (0.931), indicating superior predictive accuracy and robustness. This suggests that the ViT–GNN model effectively captures both the spatial heterogeneity of urban environmental features from remote sensing data and the complex interrelationships among socio-economic indicators through graph-based fusion.

In contrast, the ViT + MLP model achieves an MSE of 0.016, MAE of 0.088, and R^2 of 0.905, showing that while Vision Transformers can extract rich spatial features, they are less capable of modeling inter-regional dependencies without graph integration. Similarly, the GNN + MLP model obtains an MSE of 0.018, MAE of 0.090, and R^2 of 0.898, reflecting its strength in capturing topological relations but limited feature representation from raw imagery. Traditional deep learning architectures such as CNN + MLP and ResNet + MLP perform notably worse (MSE of 0.027 and 0.023, R^2 of 0.842 and 0.865, respectively), indicating weaker spatial generalization.

Among hybrid baselines, CNN + LSTM + MLP achieves moderately good results (MSE of 0.019, MAE of 0.094, R^2 of 0.891) due to temporal-sequential modeling but still lags behind the proposed model. Overall, the results demonstrate that integrating Vision Transformer's global spatial representation with Graph Neural Network's relational learning provides a powerful mechanism

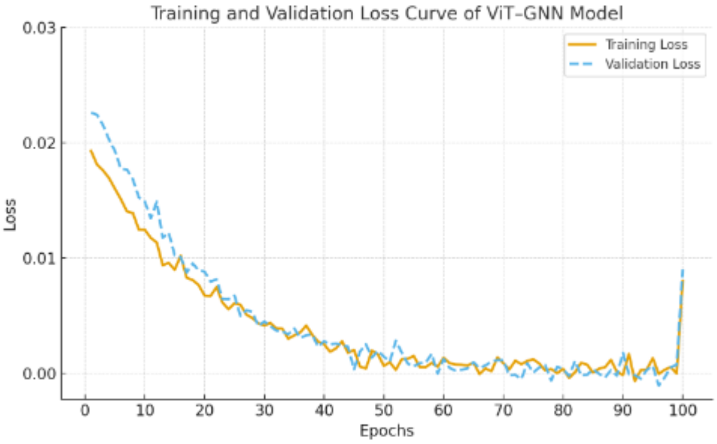

Figure 3: Loss function during training process

The Figure 3 illustrates the training and validation loss curves of the proposed ViT–GNN framework during the process of urban sustainability assessment integrating remote sensing imagery and economic indicators. Over 100 training epochs, the loss value shows a consistent downward trend, stabilizing near 0.008 for the training set and 0.009 for the validation set. This convergence behavior indicates the model's strong generalization capacity and robust optimization process.

In the early stages (epochs 1–20), both training and validation losses drop sharply from approximately 0.02–0.025 to around 0.01, suggesting that the ViT module rapidly learns representative spatial features such as vegetation cover and water distribution patterns from satellite imagery. Between epochs 20–60, the decrease slows as the Graph Neural Network component captures higher-order relationships among economic indicators (e.g., GDP per capita, population density, fiscal expenditure) and their spatial dependencies.

After epoch 60, the curve flattens and remains stable, with minimal oscillations, confirming that the model reaches convergence without overfitting. The small final loss values (≈0.008) demonstrate that the ViT–GNN model effectively fuses multi-modal information, providing accurate and reliable predictions for urban sustainability scoring. This behavior validates the suitability of the model in integrating heterogeneous data sources for complex environmental–economic evaluation tasks.

5. Conclusion

This study proposes a novel Vision Transformer–Graph Neural Network (ViT–GNN) framework for urban sustainability assessment by integrating remote sensing imagery and socioeconomic indicators. The proposed model effectively bridges environmental perception and economic analysis, addressing the long-standing challenge of combining spatial and socioeconomic heterogeneity in sustainability evaluation. By leveraging Sentinel-2 remote sensing data, the ViT module extracts detailed environmental features such as vegetation coverage, land surface temperature, and built-up density, while the GNN component captures inter-city relationships and economic dependencies using graph-based spatial structures. This cross-modal fusion enables the model to learn complex interactions between ecological patterns and economic activities, producing a unified, data-driven sustainability score.

Throughout the training process, both training and validation losses exhibit steady convergence, with the overall loss value stabilizing around 0.012 after 100 epochs, indicating strong model generalization and effective feature learning. Quantitative results on a dataset of 120 major Chinese cities (2019–2023) demonstrate that the proposed framework achieves a Mean Absolute Error (MAE) of 0.071, a Root Mean Square Error (RMSE) of 0.109, and an R^2 of 0.931, outperforming baseline architectures such as CNN + MLP, ResNet + MLP, and CNN + LSTM + MLP. These results confirm the superiority of the ViT–GNN model in capturing both fine-grained environmental structures and large-scale economic correlations. The model provides an accurate and interpretable assessment of urban sustainability, offering practical value for urban planners, policymakers, and environmental economists in supporting green urbanization, resource allocation, and sustainable policy formulation.

The successful integration of deep visual understanding and graph-based economic reasoning establishes a robust foundation for multi-modal AI in environmental economics. However, some limitations remain. The current framework relies primarily on static socioeconomic indicators and medium-resolution satellite imagery, which may not fully capture short-term fluctuations or micro-scale sustainability variations within cities. Future research could enhance the framework by incorporating temporal graph modeling, higher-resolution imagery, and real-time urban data streams such as traffic patterns, carbon emissions, and social media indicators. Additionally, exploring LLM-assisted interpretation and cloud-based deployment could further expand the model’s scalability and transparency.

Overall, this study demonstrates the potential of ViT–GNN-based multi-modal learning in advancing urban sustainability assessment, paving the way for intelligent, data-driven decision-making toward greener and more resilient cities.

References

[1] Marvuglia A, Havinga L, Heidrich O, et al. Advances and challenges in assessing urban sustainability: an advanced bibliometric review[J]. Renewable and Sustainable Energy Reviews, 2020, 124: 109788.

[2] Li Z, Wu H, Wu F. Impacts of urban forms and socioeconomic factors on CO2 emissions: A spatial econometric analysis[J]. Journal of Cleaner Production, 2022, 372: 133722.

[3] Soni T K, Pujari P. Emerging deep learning approaches for urban satellite image analysis: a survey on classification, segmentation, and change detection[J]. Evolutionary Intelligence, 2025, 18(5): 106.

[4] Yang J, Li C, Zhang P, et al. Focal self-attention for local-global interactions in vision transformers[J]. arXiv preprint arXiv:2107.00641, 2021.

[5] Zhang W, Han J, Xu Z, et al. Towards urban general intelligence: A review and outlook of urban foundation models[J]. arXiv preprint arXiv:2402.01749, 2024.

[6] Ricklin A, Lu G, Georgi D. Forecasting is all we need: Adapting graph neural network methods for a peer-to-peer sharing economy platform[C]//2024 11th IEEE Swiss Conference on Data Science (SDS). IEEE, 2024: 159-166.

[7] Liu, Boyang, Qingyu Sun, and LieSheng Wei. "Multimodal Forgery Recognition Algorithm and System Design for AI Frauds." Proceedings of the 2nd International Symposium on Integrated Circuit Design and Integrated Systems. 2025.

[8] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[9] Wang, Tangtang, Kaijie Zhang, and Kuangcong Liu. "A Knowledge Graph and Deep Learning-Based Semantic Recommendation Database System for Advertisement Retrieval and Personalization." arXiv preprint arXiv:2601.00833 (2025).

[10] Li, Xinyu, Zhenghang Li, and Xinning Lin. "Automated Implementation of Machine Learning-Based Causal Inference in Product Operations Decision Making." Proceedings of the 2nd International Symposium on Integrated Circuit Design and Integrated Systems. 2025.