Transactions in GIS

WILEY

Transactions in GIS

RESEARCH ARTICLE

Geographical Graph Attention Networks: Spatial Deep Learning Models for Spatial Prediction and Exploratory Spatial Data Analysis

Zhenzhi Jiao[1] | Ran Tao[2]

[1]The University of Hong Kong Shenzhen Institute of Research and Innovation, Shenzhen, China | [2]School of Geosciences, University of South Florida, Tampa, Florida, USA

Correspondence: Ran Tao (rtao@usf.edu)

Received: 16 October 2024 | **Revised:** 4 March 2025 | **Accepted:** 6 March 2025

Keywords: deep learning | ESDA | GeoAI | spatial prediction | spatial similarity

ABSTRACT

Some recent geospatial artificial intelligence (GeoAI) models have contributed to bridging the gap between artificial intelligence (AI) and spatial analysis. However, existing models struggle with handling small sample sizes for spatial prediction tasks across large areas. For exploratory spatial data analysis (ESDA), they are susceptible to distortion from local outliers and lack reliable interpretability methods that consider causal relationships. This study proposes Geographical Graph Attention Networks (GeoGATs), which are spatial deep learning models based on the principle of spatial (geographic) similarity. Two variants of the model are designed, namely GeoGAT-P for spatial prediction and GeoGAT-E for ESDA. Case studies using U.S. election data and homicide data demonstrate that GeoGAT-P can achieve more accurate predictions over a large spatial extent with a small sample size than existing models. GeoGAT-E can achieve decent performance in comparison with existing models and understand complex spatial relationships. Our study demonstrates how spatial similarity can be integrated with the latest deep learning models, offering valuable insights for the future direction of GeoAI research.

1 | Introduction

With the advent of multi-source big datasets, the enhancement of computational power, and advancements in artificial intelligence (AI) theories and technologies, AI has emerged as a prominent scientific field (Gao, Hu et al. 2023; Wooldridge 2021). Machine learning (ML) and deep learning (DL) models can learn non-linear relationships and features within data, effectively handling and solving various non-linear problems, thereby excelling in many complex tasks (Hastie et al. 2009; LeCun et al. 2015). ML and DL are extensively applied in spatial data prediction and classification, such as land use classification, spatiotemporal traffic flow prediction, global-scale ecological mapping, and so on (Alshari et al. 2023; Christin et al. 2019; Zhang et al. 2019). However, directly applying ML and DL techniques to spatial analysis faces challenges due to the characteristics of spatial data (Meyer et al. 2018; Nikparvar and Thill 2021). For example, most ML models are based on the assumption of independent and identically distributed (i.i.d.) data, but the inherent spatial autocorrelation in spatial data can cause the models to be overoptimistic (Meyer and Pebesma 2022; Meyer et al. 2019; Ploton et al. 2020). Moreover, the spatial heterogeneity in spatial data makes it difficult for standard DL models to adapt to the characteristics of all regions (Li et al. 2023; Yuan et al. 2020).

Spatial prediction and exploratory spatial data analysis (ESDA) are fundamental components of spatial analysis, focusing on predicting and understanding the complex relationships and patterns within spatial data (Anselin 2002; Harris et al. 2010; Ward and Gleditsch 2018). Geospatial Artificial Intelligence (GeoAI) has become an important subfield within GIScience that considers the characteristics of spatial data to bridge the gap between spatial analysis and AI (Janowicz et al. 2020; Liu and Biljecki 2022). Some research efforts have been dedicated to integrating AI tools with spatial analysis to better address the tasks of spatial prediction and ESDA. For example, Georganos et al. (2021) proposed geographical Random Forests (GRF), which calibrate the Random Forest model by adopting the principles of Geographically Weighted Regression (GWR) through the establishment of local models. Du et al. (2020) developed a Geographically Neural Network Weighted Regression (GNNWR) model that combines Ordinary Least Squares (OLS) and neural networks to predict unknown observations and estimate spatial heterogeneity. Zhu et al. (2022) proposed Spatial Regression Graph Convolutional Neural Networks (SRGCNNs), employing the conception of a geographically weighted method to consider the spatial relationships of observations for spatial prediction and regression.

The aforementioned models are calibrated to account for spatial dependence and spatial heterogeneity, significantly advancing the integration of AI with spatial analysis. However, several challenges remain in spatial prediction and ESDA. For spatial prediction, models based on spatial dependence and heterogeneity often rely on dense observations to establish the stationarity assumption or to construct local models effectively. This reliance makes them to hardly to achieve accurate spatial predictions across large areas with a small sample size. For ESDA, spatial analysis based on spatial dependence and spatial heterogeneity is prone to being influenced by local outliers, which can distort local relationships and fail to provide convincing results. Furthermore, some models provide interpretable methods to help improve understanding of complex spatial patterns, but traditional multi-layer perceptrons[1] (MLP) directly learn the statistical correlations between input features and output labels during training without distinguishing between causal and non-causal features (Goodfellow et al. 2016; Louizos et al. 2017; Zhang et al. 2021). For ESDA, this process may absorb some shortcut features, leading to misunderstandings of spatial patterns and confounding effects.

In recent years, graph-based DL has gained widespread application in geospatial data science due to its capability to handle heterogeneous nodes in graph-structured data, which offers a foundational method to address the above issues (Xu and Zuo 2024; Zheng and Lu 2024; Zhu et al. 2022). Specifically, the Graph Attention Networks (GAT) model is a neural network architecture based on a non-i.i.d. assumption that operates on graph-structured data, leveraging an attention mechanism to effectively capture the varying influences of different neighbors on the target node (Brynte et al. 2024; Vrahatis et al. 2024; Xie et al. 2024). In this study, we propose the Geographical Graph Attention Networks (GeoGATs) for spatial prediction and ESDA, called GeoGAT-P and GeoGAT-E, respectively. GeoGAT-P employs a GAT model calibrated by the principle of spatial similarity, which relaxes the constraints of spatial dependence and spatial heterogeneity to enhance spatial prediction capabilities across large areas. GeoGAT-E integrates spatial heterogeneity and spatial similarity to build the local model, which can avoid the influence of local outliers in ESDA. Additionally, we introduce the causal attention mechanism and backdoor adjustment to capture spatial relationships and map the local feature importance score for interpretability. The remainder of this paper is structured as follows. Section 2 represents the GeoGAT models; Section 3 contains datasets and data preprocessing; Section 4 presents the results of GeoGAT-P; Section 5 shows GeoGAT-E, as well as the local feature importance map (LFIM); Section 6 discusses our insights and limitations of GeoGAT models; Section 7 displays the conclusions.

2 | Geographical Graph Attention Networks

2.1 | Spatial Similarity

Spatial similarity, as summarized by Zhu et al. (2018), refers to "the more similar the geographical configurations of two spatial units, the more similar the values of the target variable at these two spatial units" (Zhu et al. 2018; Zhu and Turner 2022). For example, in the monitoring of heavy metal pollution, the levels of heavy metal contamination are often closely associated with multiple factors, such as geological background, soil properties (e.g., pH value and organic matter content), hydrological conditions, types and scales of industrial activities, land use patterns, and waste disposal methods. If two spatial units are more similar in these geographical and human activity characteristics, their levels of heavy metal pollution tend to be closer as well.

Given a spatial unit, the geographical configuration can be represented as a vector:

$$X = \left(x_1, x_2, \ \dots \ , x_n\right) \tag{1}$$

where n is the number of input features.

The value of spatial similarity between spatial units i and j can be determined as:

$$S_l(i,j) = \sum_{v=1}^{n} P\{C(i,j)\} \tag{2}$$

$S_l(i,j)$ is the local spatial similarity between the spatial unit i and j; $P(\cdot)$ is the function of weighted average method or average method, and weighted average method can use the feature importance of each variable calculated by Random Forest (Zhu et al. 2015), this study used average method; the $C(\cdot)$ function is employed to determine the spatial similarity at the vth feature x_v, which is calculated as:

$$C_v(i,j) = \exp\left(- \frac{\left(x_v(i) - x_v(j)\right)^2}{2\sigma} \right) \tag{3}$$

$x_v(i)$ and $x_v(j)$ are the values of spatial units i and j at feature v; σ is the deviation of feature x_v.

According to Zhu et al. (2018), spatial similarity can consider not only the geographic configuration similarity between two spatial units but also the similarity in the composition and structure of geographic variables within the spatial neighborhoods surrounding these units. To this end, we further extend spatial similarity to incorporate the similarity of the neighborhoods of spatial units (Figure 1).

The formula for the similarity of neighborhoods is defined as follows:

$$S_n(N(i), N(j)) = \sum_{v=1}^{n} P\{C(N(i), N(j))\} \tag{4}$$

$S_n(N(i), N(j))$ is the spatial similarity value between the neighborhoods of spatial units i and the neighborhoods of spatial unit j; the $C(\cdot)$ function follows Equation (3), which is used to calculate the spatial similarity at the vth feature, using mean value of two neighborhoods, $\bar{x}_v$. In this way, we define the overall spatial similarity of two spatial units as follows:

$$S(i,j) = S_l(i,j) + \alpha \times S_n(N(i), N(\text{j})) \tag{5}$$

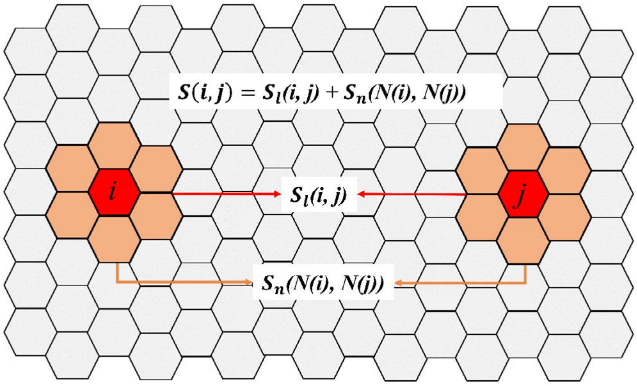

FIGURE 1 | The framework of spatial similarity used in GeoGAT-P.

$S(i,j)$ is the spatial similarity between spatial unit i and spatial unit j; α is the coefficient, this study set 1 (see Appendix A for details).

2.2 | GeoGATs for Prediction

As shown in Figure 2, we first calculate the spatial similarity between a spatial observation and other observations, extracting the 10 most similar observations (the selection process is detailed in Appendix B) as neighboring nodes. Next, the spatial similarities are set as edge weights to help the model better understand spatial relationships. A masking operation ensures that each node only exchanges information and aggregates features with its neighboring nodes (color differentiation aims to present a global subgraph for each observation that includes unknown and known observations). In the training process, the dropout mask would randomly drop some parameters in GeoGAT-P to avoid overfitting. Finally, the model outputs the predicted values and generates a residual map.

GAT is a neural network model designed for graph data that introduces an attention mechanism to handle graph-structured data, enabling flexible analysis of relationships between nodes (Velickovic et al. 2017). A spatial dataset is modeled as a graph $G=(V,E)$, with V denoting the nodes in the graph. Each node $v \in V$ is typically associated with a vector h_v. E represents the set of edges, which indicate the connections between the nodes in the graph. Each edge $E(i,j) \in E$ links node i and node j, and each edge can have a weight associated with it.

The spatial similarity matrix S is used as a mask operation to determine the adjacent matrix of GAT. The formulation is as follows:

$$M_{ij} = S = \begin{cases} 1, & \text{if } j \in N(i) \\ 0, & \text{else} \end{cases} \tag{6}$$

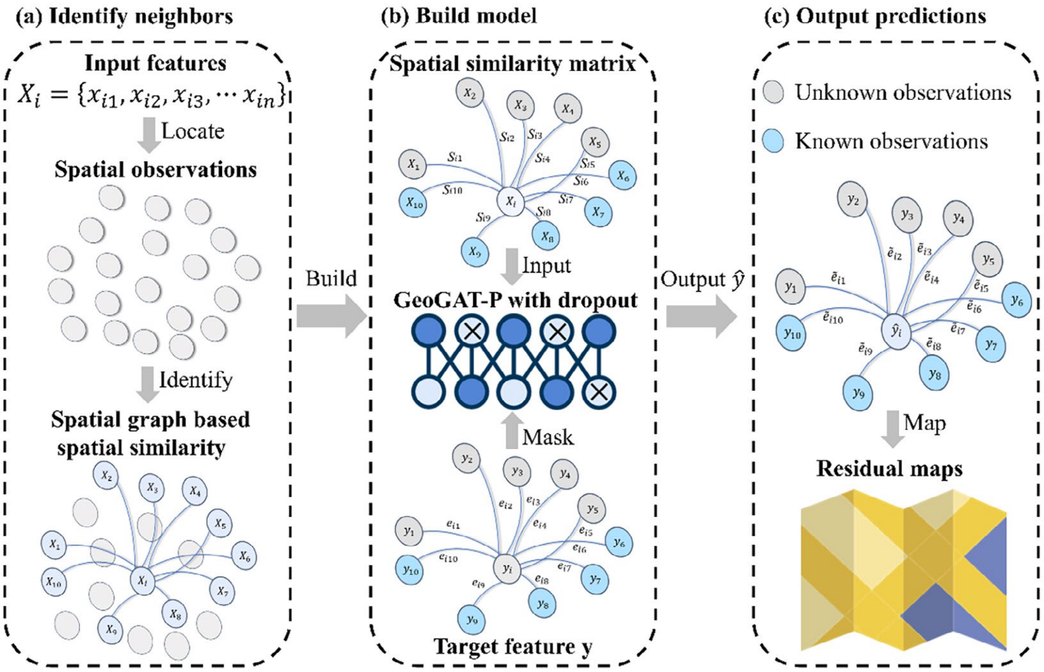
(a) Identify neighbors
Input features
$X_i = \{x_{i1}, x_{i2}, x_{i3}, \cdots x_{in}\}$
Locate
Spatial observations
Identify
Spatial graph based spatial similarity
X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10} X_i
Build
(b) Build model
Spatial similarity matrix
S_{i1} S_{i2} S_{i3} S_{i4} S_{i5} S_{i6} S_{i7} S_{i8} S_{i9} S_{i10}
Input
GeoGAT-P with dropout
Mask
y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8 y_9 y_{10} y_i
e_{i1} e_{i2} e_{i3} e_{i4} e_{i5} e_{i6} e_{i7} e_{i8} e_{i9} e_{i10}
Target feature y
Output $\hat{y}$
(c) Output predictions
Unknown observations
Known observations
$\hat{y}_i$
$\tilde{e}_{i1}$ $\tilde{e}_{i2}$ $\tilde{e}_{i3}$ $\tilde{e}_{i4}$ $\tilde{e}_{i5}$ $\tilde{e}_{i6}$ $\tilde{e}_{i7}$ $\tilde{e}_{i8}$ $\tilde{e}_{i9}$ $\tilde{e}_{i10}$
Map
Residual maps

FIGURE 2 | The overall workflow of GeoGAT-P for spatial prediction.

M_{ij} represents the connections between each node i and its 10 most similar node j; $N(i)$ denotes the neighbors of node i^2.

By establishing graph relationships, GAT uses a linear transformation on node features:

$$h'_j = W h_j \tag{7}$$

where W is the weight matrix to be learned, and h_j is the input feature of the node j.

GAT calculates the attention scores between the target node and its neighboring nodes through a shared attention mechanism. The initial attention scores e_{ij} are computed using a shared attention mechanism. The formula is:

$$e_{ij} = \text{LeakyReLU}\left(\alpha^T \left[W h_i \parallel W h_j\right]\right) \tag{8}$$

where e_{ij} represents the initial attention score between node i and node j; LeakyReLU is the Leaky Rectified Linear Unit activation function; α is a learnable weight vector; W is a learnable weight matrix; h_i and h_i are the feature vectors of node i and node j, respectively; $\parallel$ denotes the concatenation operation.

Using the attention mechanism, the values of spatial similarity are applied as edge weights, which are defined as $S(i,j)$, Equation (5). The initial attention scores are then combined with the pre-computed edge weights S_{ij}:

$$\widetilde{e}_{ij} = S_{ij} \cdot e_{ij} \tag{9}$$

where $\widetilde{e}_{ij}$ represents the adjusted attention weight between node i and node j.

The final attention weights α_{ij} are obtained by normalizing the adjusted attention scores with the softmax function:

$$\alpha_{ij} = \frac{\exp\left(\widetilde{e}_{ij}\right)}{\sum_{k \in N(i)} \exp\left(\widetilde{e}_{ik}\right)} \tag{10}$$

where α_{ij} represents the final attention weight between node i and node j; $N(i)$ denotes the set of neighbors of node i.

The weighted aggregation of the neighbor features is carried out employing the final attention weights:

$$h''_i = \sigma\left(\sum_{j \in N(j)} \alpha_{ij} W h_j\right) \tag{11}$$

h''_i represents the updated feature vector of a node i after aggregating the features of its neighbors; σ is an activation function, ReLU. The prediction formula can be written as:

$$\widehat{y}_i = \emptyset\left(h'_i\right) \tag{12}$$

where $\emptyset$ can be a linear transformation of the output layer.

The model adopts Mean Squared Error (MSE) for the loss function to compute gradients and applies the Adam optimization algorithm for updating its parameters. MSE is employed to quantify the discrepancy between the model's predicted values and the actual values. The formula is given by:

$$L_{\text{mse}} = \frac{1}{n} \sum_{i \in n} \left(\widehat{y}_i - y_i\right)^2 \tag{13}$$

where L_{mse} is the loss function value; $\widehat{y}_i$ is the predicted value; y_i is the true value; n is the number of train dataset.

We used Bayesian optimization to find the network hyperparameters (including the dropout rate and number of layers) and achieve optimal model performance. The model employs Dropout and L2 as regularization techniques to prevent overfitting (Baldi and Sadowski 2013; Cortes et al. 2012). During the forward pass, Dropout is applied to the input features and the output of the first graph attention convolutional layer. The fraction of neurons to drop is specified by the dropout rate p. The formulation is as follows:

$$x' = \frac{1}{1-p} m \odot x \tag{14}$$

x is the input tensor; the binary mask vector m is drawn from a Bernoulli distribution with probability p; $\odot$ represents element-wise multiplication; x' is the output tensor after applying dropout; p is the dropout rate; in this study, we set it to 0.2.

L2 regularization can prevent the parameters from becoming too large, which can lead to overfitting. The L2 regularization term is as follows:

$$L_{L2} = \frac{\lambda}{2} \sum_j \theta^2{}_j \tag{15}$$

where L_{L2} is the L2 regularization term; λ is the regularization strength (weight decay coefficient); θ_j are the model parameters.

Gradients are computed based on the MSE loss function, represented as:

$$g_t = \nabla_\theta L\left(\theta_{t-1}\right) \tag{16}$$

where g_t represents the gradient of the loss function concerning the model parameters at time step t; θ_{t-1} are the model parameters at the previous time step; ∇_θ denotes the gradient with respect to the model parameters θ; L is defined as:

$$L = L_{\text{mse}} + L_{L2} \tag{17}$$

The Adam optimization algorithm updates model parameters using the calculated gradients. Update rules for the Adam optimizer include first-order moment estimation, second-order moment estimation, and bias correction. The final parameter update formula is:

$$\theta_t = \theta_{t-1} - \eta \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t + \epsilon}} \tag{18}$$

where θ_t are the model parameters at time step t; η represents the learning rate; $\hat{m}_t$ denotes the bias-corrected first-order moment estimate; $\hat{v}_t$ signifies the bias-corrected second-order moment estimate; ϵ is a small constant.

In this study, we employed MSE and Mean Absolute Error (MAE) to evaluate the performance of models on prediction. MSE is represented in Equation (13). The formulation of MAE is as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{19}$$

where n is the number of test datasets.

2.3 | GeoGATs for ESDA

In Figure 3, spatial data is the first input to identify neighboring nodes using optimal bandwidth, and then spatial similarity is calculated to determine the spatial similarity matrix as edge weights to establish a local model. Second, we design a causal attention mechanism and backdoor adjustment to deal with causal features and shortcut features. G represents graph-structured data; C are causal features, which are the predictors that truly affect the target feature y; S represents confounding variables, which are biases and noise in the data. Third, a masking operation facilitates information exchange and feature aggregation between nodes to train the model. Finally, the high-dimensional features extracted by the graph convolution layers are normalized and projected back to the original input feature space using linear layers to generate the map.

The kernel functions determine the spatial weights of neighbors that decay based on distance for GWR (Brunsdon et al. 1996). For instance, the Gaussian kernel function can be expressed as:

$$W_{ij} = \exp\left(-\frac{d^2{}_{ij}}{2b^2}\right) \tag{20}$$

where d_{ij} denotes the distance between observations i and j, and b is the bandwidth parameter.

GeoGAT-E references the analysis approach of GWR, using the Akaike Information Criterion (AIC) to determine the optimal bandwidth, which defines the neighbors for each local observation[2] (Li et al. 2020). The optimal bandwidth represents the best spatial scale for the model to capture local spatial characteristics, enabling the model to adequately capture spatial heterogeneity (Fotheringham et al. 2022). However, unlike GWR, which uses spatial weights to determine neighbor contributions, the GeoGAT-E model calculates the weights $W_S(i,j)$ for local fitting based on spatial similarity, which enables ESDA to avoid the influence from local outliers. GeoGAT-E establishes local models through graph connection operations. This approach ensures that the local relationships between nodes are relatively robust, allowing the use of the similarity between two nodes directly without considering their neighborhoods. The formulation is as follows:

$$W_s(i,j) = S_l(i,j)\cdot e_{ij} \tag{21}$$

GeoGAT-E employs spatial adjacency matrix W as an adjacency mask to define the adjacency matrix:

$$M_{ij} = W = \begin{cases} 1, & \text{if } j \in N(i) \\ 0, & \text{else} \end{cases} \tag{22}$$

The training process of GeoGAT-E is similar to the GeoGAT-P, which can be seen in Equations (13–18). $N(i)$ denotes the neighbors of node i. The purpose of GeoGAT-E is ESDA, which requires performing in-sample predictions on the entire dataset to ensure that the model fully understands the data. Dropout parameters are not necessary for DL models to predict in-sample dataset, as confirmed by a recent study (Özgür and Nar 2020).

Recently, Sui et al. (2022) proposed the causal attention learning strategy, which aims to identify spatial patterns and mitigate

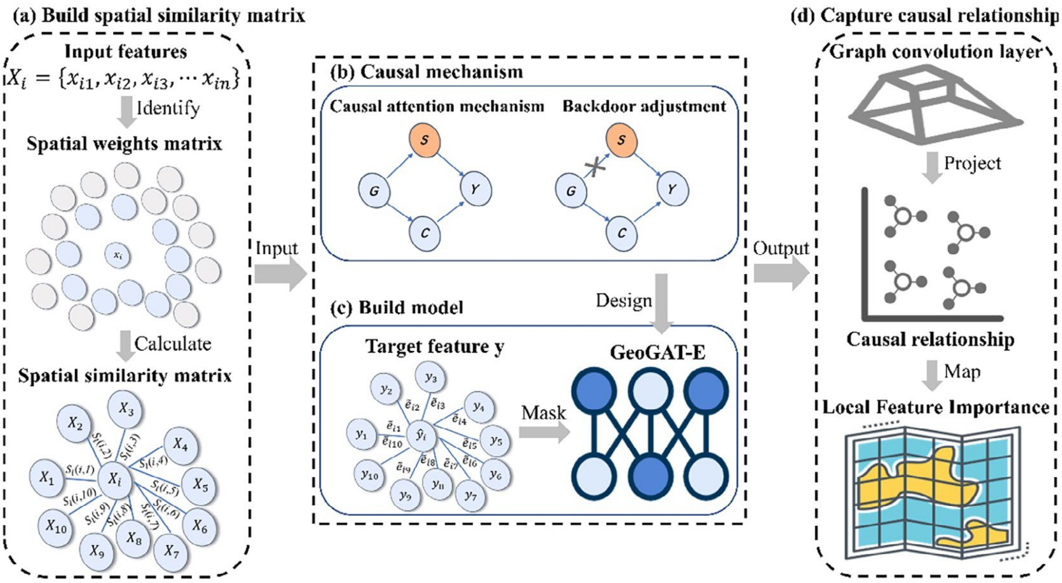
(a) Build spatial similarity matrix
Input features
$X_i = \{x_{i1}, x_{i2}, x_{i3}, \cdots x_{in}\}$
Identify
Spatial weights matrix
x_i
Calculate
Spatial similarity matrix
X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10} X_i
$S_{(i,1)}$ $S_{(i,2)}$ $S_{(i,3)}$ $S_{(i,4)}$ $S_{(i,5)}$ $S_{(i,6)}$ $S_{(i,7)}$ $S_{(i,8)}$ $S_{(i,9)}$ $S_{(i,10)}$
Input
(b) Causal mechanism
Causal attention mechanism
Backdoor adjustment
S G Y C
(c) Build model
Design
Target feature y
y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8 y_9 y_{10} y_{11} $\hat{y}_i$
Mask
GeoGAT-E
Output
(d) Capture causal relationship
Graph convolution layer
Project
Causal relationship
Map
Local Feature Importance

FIGURE 3 | The overall workflow of GeoGAT-E.

the effects of shortcut confounding, thereby enhancing both the interpretability and generalizability in graph classification (Sui et al. 2022). Furthermore, Zhao and Zhang (2024) developed a spatial–temporal explanation model for Dynamic Graph Neural Networks based on causal theory, aiming to enhance the interpretability and generalization capability of models dealing with spatiotemporal structure data (Zhao and Zhang 2024). Inspired by both studies, we utilize two parallel GAT layers to extract causal and shortcut features from the input data. The causal GAT layer focuses on identifying features with direct causal influence on the target feature, while the shortcut GAT layer captures spurious correlations. The extracted features are then normalized using batch normalization.

$$C = \mathrm{ELU}\left(\mathrm{GATConv}_c\,(X, A, E)\right) \tag{23}$$

$$S = \mathrm{ELU}\left(\mathrm{GATConv}_s\,(X, A, E)\right) \tag{24}$$

where X is the input feature matrix; ELU (exponential linear unit) is an activation function; A is the adjacency matrix; and E is the edge weight matrix; GATConv is the graph attention convolution; C and S are the causal features and confounding features, respectively.

This study extracts the interpretability of causal features through a GATConv in causal attention mechanism using linear layers. The core mechanism in GATConv is the attention weights, which determine the contribution of each neighboring node's features during aggregation (Vaswani 2017; Lin et al. 2017). Through these weights, the model can capture more complex node relationships and patterns, particularly in scenarios where causal relationships or feature contributions are clear. (Wiegreffe and Pinter 2019). To improve model interpretability, a linear layer is used to map the high-dimensional feature vectors obtained through the attention mechanism back to the original input feature dimensions. This mapping makes the relationship between each input feature and its representation in the high-dimensional embedding space clearer, thereby making the model's output easier to interpret. The formula is as follows:

$$C_{\mathrm{mean}} = \mathrm{ReLU}\left(W_c C + b_c\right) \tag{25}$$

$$S_{\mathrm{mean}} = \mathrm{ReLU}\left(W_s S + b_s\right) \tag{26}$$

W_c and W_s are the weight matrices of the linear layers. b_c and b_s are the bias terms. C_{mean} and S_{mean} are the projected causal and confounding features with the same dimensions as the input features.

By introducing a causal intervention loss function, the model can supervise causal features and confounding features separately. Independent weight updates and different supervision signals during the training process enable the model to learn different types of features. The causal intervention loss function is defined as follows:

$$L\text{causal intervention} = \mathrm{MSE}\left(C_{\mathrm{mean}}, y\right) + \mathrm{MSE}\left(S_{\mathrm{mean}}, y\right) \tag{27}$$

$MSE(\cdot)$ can be seen in Equation (13). Thus, the total loss of the model is:

$$L = \mathrm{LGAT} + \lambda \ldots L\text{causal intervention} \tag{28}$$

λ is the weight of the causal intervention loss; we set it as 0.5 in this paper.

The backdoor adjustment is an effective causal inference method that helps a model accurately estimate causal effects in the presence of confounding variables. In graph neural networks, implementing backdoor adjustment through a causal intervention loss function aids in separating causal features from shortcut features. The adjustment formula is specifically implemented as:

$$P(y|\,do(C)) = \sum_{s} P(y|\,C, S)P(S) \tag{29}$$

Global feature importance is calculated by summing the absolute values of the local feature importance scores of each feature across all samples. The formula is as follows:

$$I_j = \sum_{i=1}^{N} \left| C_{\mathrm{mean}(i,j)} \right| \tag{30}$$

$C_{\mathrm{mean}(i,j)}$ is the local feature importance score of the jth feature in the ith sample; N is the number of samples; j is the index of the feature.

For ESDA, we selected Mean Absolute Percentage Error (MAPE) and the determination coefficient (R^2) as metrics to describe the model's goodness of fit. The formulations are defined as:

$$\mathrm{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_y}{y_i} \right| \times 100\,\% \tag{31}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} \tag{32}$$

3 | Datasets and Data Preprocessing

This study adopts the public datasets of U.S. Elections and Homicides for empirical case studies, both of which were downloaded from GeoDa Lab's website (https://geodacenter.github.io/data-and-lab/). Due to the presence of global outliers in the Homicides dataset, we used three standard deviations to exclude them (Huber 2018). This paper used Pearson correlation coefficients to exclude variables with correlations greater than 0.7 (Hulland 1999). Tables 1 and 2 list the datasets' descriptive statistics. Dependent variables (pct_gop 16 and HR90, respectively) of the datasets are illustrated in Figures 4 and 5.

4 | Spatial Prediction

The datasets were divided into a 10% training set, a 10% validation set, and an 80% test set. This strategy helps us simulate spatial prediction in data-sparse situations across large areas.

TABLE 1 | Description of the U.S. Elections dataset.

Variables	Min	Max	Mean	Std.	Description
AGE_U5	0	13.3	5.8841	1.1882	Persons under 5 years, percent, 2014
AGE_O65	0	52.9	17.6353	4.401	Persons 65 years and over, percent, 2014
LO5	0	95.6	9.1224	11.3976	Language other than English spoken at home, pct age 5+, 2009–2013
LH1	50.8	99.8	86.432	4.3942	Living in same house 1 year & over, percent, 2009–2013
Female_pct	0	56.8	49.9387	2.3806	Female persons, percent, 2014
White_pct	0	99.3	85.454	15.7397	White alone, percent, 2014
EDU_H25	45	99	84.5088	6.9124	High school graduate or higher, percent of persons age 25+, 2009–2013
EDU_B25	3.2	74.4	19.7357	8.8315	Bachelor's degree or higher, percent of persons age 25+, 2009–2013
MTW_16	8.2	44.2	23.0892	5.361	Mean travel time to work (minutes), workers age 16+, 2009–2013
HR	19.4	93.8	72.2906	7.8606	Homeownership rate, 2009–2013
pct_gop 16	0.0412	0.9527	0.6366	0.156	Votes for Republican candidate as percent of total votes, 2016

TABLE 2 | Description of the U.S. Homicides dataset.

Variables	Min	Max	Mean	Std.	Description
RD90	−2.4103	5.5831	−0.0246	0.9663	Resource deprivation in the year, 1990
PS90	−3.8351	3.6022	−0.0025	0.9857	Population structure, 1990
UE90	0	30.5340	6.6054	3.0126	Unemployment rate in the year, 1990
DV90	2.0755	17.1900	7.1571	1.7300	Divorce rate of the year, 1990
MA90	20	55.4	34.4334	3.5839	Median age in the years, 1990
BLK90	0	86.2360	8.2547	13.7679	Percentage of black population for the year, 1990
HR90	0	26.1030	5.8199	5.7158	Homicides rate per 100,000 (1990)

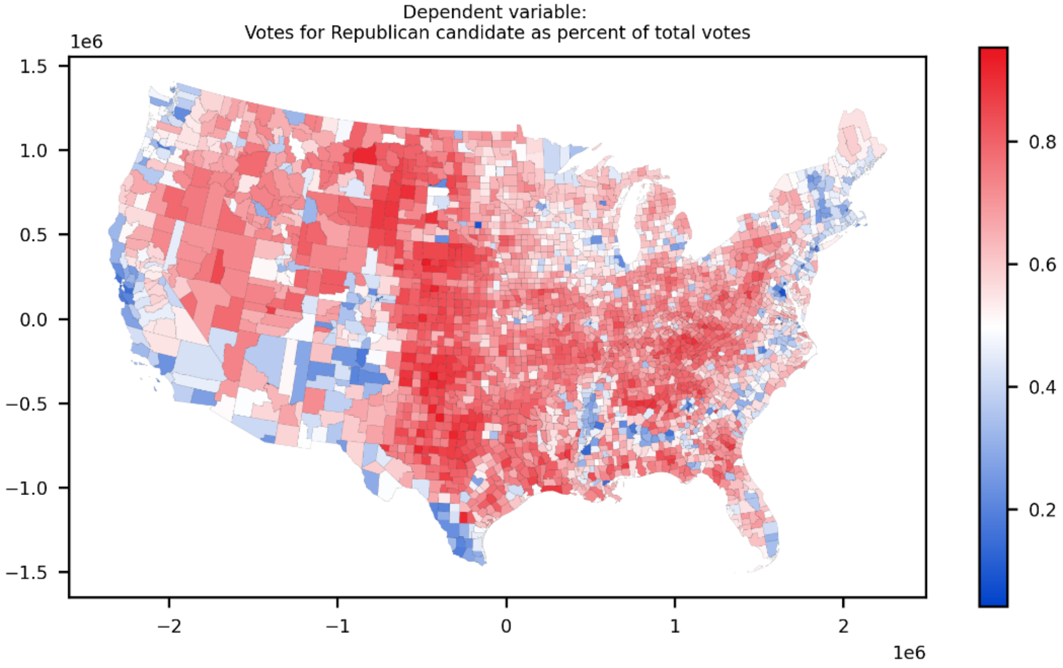

FIGURE 4 | Map of the dependent variable in the U.S. Elections dataset.

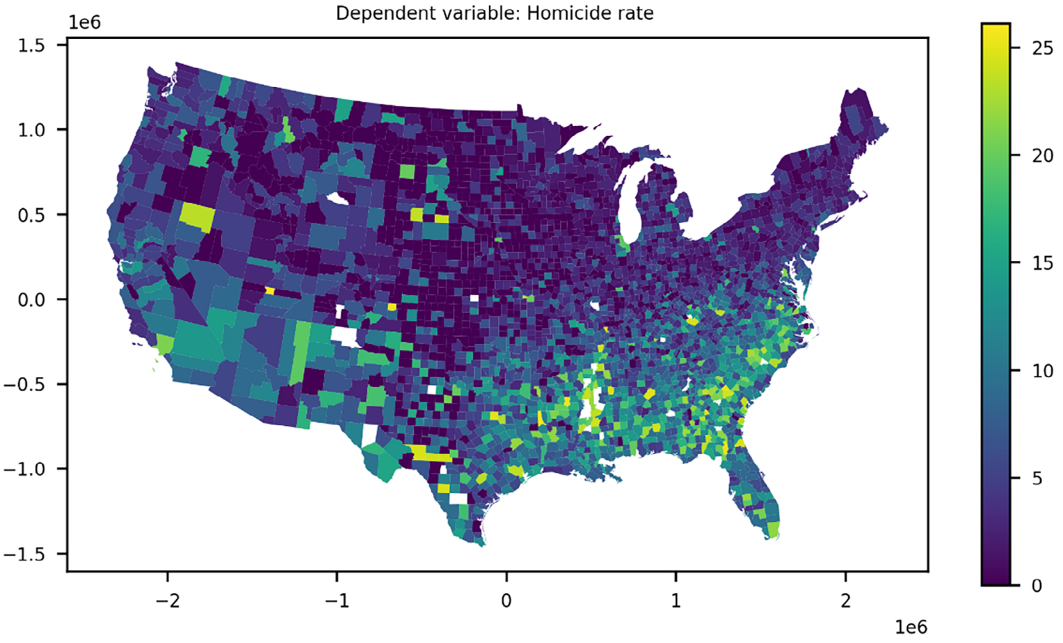

FIGURE 5 | Map of the dependent variable in the U.S. Homicides dataset.

TABLE 3 | Results of spatial predictions.

	Elections		**Homicides**	
Model	**MSE**	**MAE**	**MSE**	**MAE**
GNNWR	0.0427	0.1635	0.7564	0.5883
SRGCNN	0.0240	0.1196	0.9529	0.6659
SRGAT	0.0228	0.1974	0.9321	0.7620
GeoGAT-P*	0.0107	0.0810	0.5889	0.5131
GeoGAT-P	0.0106	0.0793	0.5385	0.4835

To avoid potential bias from random sample partitioning, we use the averages of 10 runs for each model as the evaluation metrics. In Table 3, we compared the performance of the GNNWR (https://figshare.com/s/84b78c21262dc185d44a), SRGCNN,[3] and GeoGAT-P models on the Elections dataset and the Homicides dataset. GeoGAT-P demonstrated the best predictive performance on both datasets, with the lowest MSE and MAE. The performance of SRGCNN and GNNWR was comparable; SRGCNN outperformed GNNWR on the Elections dataset, while on the Homicides dataset, SRGCNN had a lower MSE but a higher MAE than GNNWR. GeoGAT-P and SRGCNN are both graph DL models; GeoGAT-P considers spatial similarity, while SRGCNN considers spatial dependence. Compared to SRGCNN, GeoGAT-P reduced MSE by 55% and 43%, and MAE by 33% and 27% in the Elections dataset and Homicides dataset, respectively. The residual maps are shown in Figure 6.

To further understand GeoGAT-P, it is necessary to isolate the impact of spatial similarity on the GAT mechanism. Since GAT requires a clearly defined graph structure of observations, it cannot directly predict unknown observations based on known ones. To address this, we developed a baseline based on the SRGCNN, referred to as SRGAT[4]. Both SRGCNN and GeoGAT-P belong to the category of graph-based DL models, and SRGCNN is built on spatial dependence; comparing GeoGAT-P with SRGAT highlights the role of spatial similarity more effectively. From Table 3, we can see that SRGAT does not show significant improvement compared to SRGCNN. Therefore, we can conclude that the superior spatial prediction performance of GeoGAT-P is a main result of applying the spatial similarity principle rather than the influence of the GAT mechanism.

Traditionally, spatial similarity only accounts for the similarity between two observations (Zhu et al. 2015), whereas GeoGAT-P takes one step further by also considering the spatial similarity between the neighborhoods of two observations. To evaluate GeoGAT-P's advancement in spatial similarity, we also compared it with a method that considers only the similarity between two observations with a GAT model named GeoGAT-P*. Table 3 presents a performance comparison between the two, demonstrating that GeoGAT-P enhances model robustness and predictive performance by integrating both the spatial similarity between observations and the similarity between their neighborhoods.

We evaluated the model's sensitivity to different training set sizes, while keeping the size of the validation set unchanged. The predictive accuracy for both datasets is shown in Table 4. In the Elections dataset, the predictive performance of GeoGAT-P improved with the increase in training dataset size; however, the differences in predictive accuracy were not substantial. On the Homicides dataset, the model's performance showed a gradual improvement from 10% to 40% training set sizes, despite the performance being suboptimal at the 50% sampling rate. The reason may be the small size of the validation set, but we did not adjust it to ensure comparability. This indicates that the GeoGAT-P model is not sensitive to data size and still demonstrates strong predictive capabilities even at low sampling rates. Therefore, we can conclude that the model is capable of handling predictions with low sampling rates in large areas. At all sampling rates,

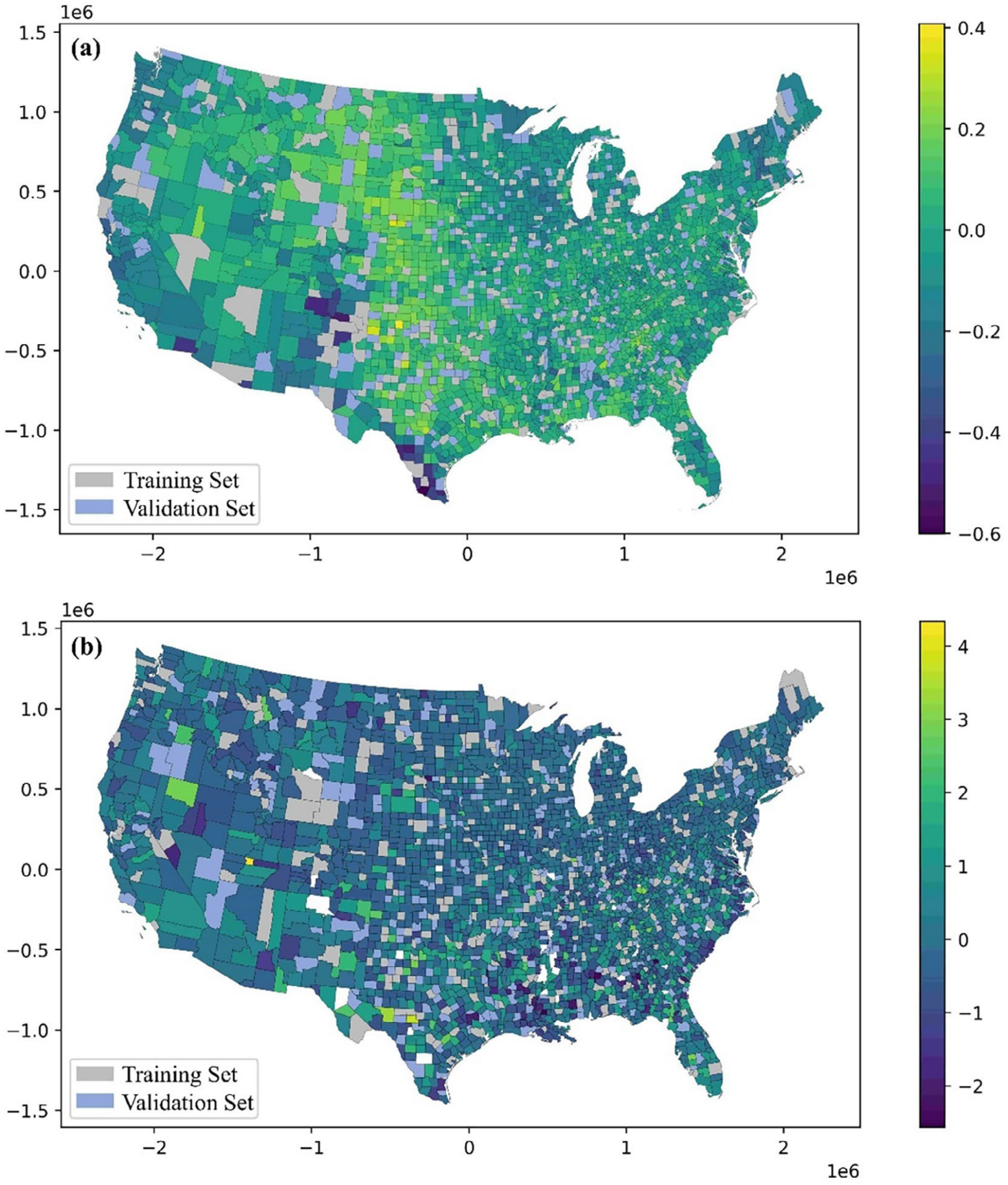

FIGURE 6 | Residual maps. (a) Residual map of the Elections dataset. (b) Residual map of the Homicides dataset.

GeoGAT-P outperformed SRGCNN and GNNWR, exhibiting smaller prediction errors. Noticeably, all models run in the following environment: the operating system is Ubuntu 22.04 LTS, the GPU is GeForce RTX 4090, the CPU is Intel Core i7, and the RAM is 32 GB. The computational efficiency of GeoGAT-P, compared to that of SRGCNN and GNNWR, is presented in Table C1 in Appendix C.

5 | Exploratory Spatial Data Analysis (ESDA)

5.1 | Results

To demonstrate the performance of the GeoGAT-E model, we compared it with Multiscale geographically Weighted Regression (MGWR) (https://mgwr.readthedocs.io/en/latest/) (Oshan et al. 2019), GRF (https://search.r-project.org/CRAN/refmans/SpatialML/html/grf.html), and SRGCNN. As shown in Table 5, GeoGAT-E achieves the best performance in both datasets, with the highest R^2 and the lowest MAPE. It is followed by GRF, which performs slightly better than MGWR in the Elections dataset and outperforms SRGCNN in the Homicides dataset. SRGCNN shows the poorest performance in the Elections dataset, while MGWR performs the worst in the Homicides dataset. Due to the presence of zero values in the dependent variable homicide rate of the Homicides dataset, MAPE is highly sensitive to zero and is thus not suitable for evaluating models' performance on the Homicides dataset. A comparison of the computational efficiency among GeoGAT-E, SRGCNN, GRF, and MGWR is provided in Table C2 (Appendix C). In Table 5, MGWR performs particularly poorly in the Homicides dataset. As shown in Figure 5, the dependent variable in the Homicides

TABLE 4 | Model predictive performances on different sampling sizes.

Ratio			10%	20%	30%	40%	50%
GeoGAT-P	Elections	MSE	0.0106	0.0102	0.0098	0.0092	0.0092
		MAE	0.0793	0.0769	0.0755	0.0731	0.0733
	Homicides	MSE	0.5385	0.5149	0.5142	0.4909	0.4415
		MAE	0.4835	0.4650	0.4776	0.4727	0.4683
SRGCNN	Elections	MSE	0.0240	0.0283	0.0231	0.0225	0.0240
		MAE	0.1196	0.1248	0.1206	0.1176	0.1186
	Homicides	MSE	0.9529	0.8614	0.8314	0.8110	0.7320
		MAE	0.6659	0.6449	0.6344	0.6327	0.6161
GNNWR	Elections	MSE	0.0427	0.0345	0.0341	0.0393	0.0372
		MAE	0.1635	0.1472	0.1469	0.1531	0.1506
	Homicides	MSE	0.6328	0.6451	0.5531	0.6202	0.6380
		MAE	0.5269	0.5321	0.5074	0.5338	0.5264

TABLE 5 | Results on ESDA.

	Elections		Homicides
Models	R^2	**MAPE (%)**	R^2
MGWR	0.9350	5.4712	0.5690
GRF	0.9411	5.1345	0.9128
SRGCNN	0.7702	7.7110	0.8625
GeoGAT-E	0.96214	3.5946	0.9841

dataset contains numerous local outliers, indicating the strong spatial non-linear effects present in the dataset. Local Indicators of Spatial Association (LISA) statistics indicate that local outliers account for 17.31% of the entire dataset (Anselin 1995). As a locally linear model, MGWR fails to address the challenges posed by these local outliers. In contrast, GeoGAT-E, GRF, and SRGCNN, all of which are non-linear models, significantly outperform MGWR in the Homicides dataset. Among them, GeoGAT-E achieves the best performance, with an R^2 value 7.4% higher than the second-best model, GRF. This demonstrates that GeoGAT-E effectively addresses the impact of local outliers by incorporating spatial similarity as a weighting mechanism.

5.2 | Model Interpretation

In the causal attention mechanism, attention scores are used to measure the importance of the impact of each input feature on the target feature. The higher the importance score, the stronger the relationship of that feature with the target feature. LFIM provides the relationship for each local observation, which helps geographers better understand spatial effects and spatial relationships. The LFIMs of two cases can be found in Figures 7 and 8. By averaging the absolute values of the local feature importance scores of each feature across all samples, we provided the global importance of each feature on the target feature. As shown in Figure 9, the empirical analysis case of the Elections dataset indicates that White_pct has the largest global contribution to the model, followed by Female_pct and MTW_16, with HR contributing the least. In the Homicides case, BLK90 contributes the most to the model, followed by UE90 and PS90, with MA90 contributing the least (Figure 10).

Furthermore, Tables 6 and 7 present the important distributions of GeoGAT-E and MGWR, respectively. Due to space constraints, we just presented the comparison in the Homicides dataset. The contribution of each variable to the model can be assessed by calculating the mean of the absolute coefficients from MGWR (Zhao et al. 2023). BLK90 and MA90 are the most and least influential variables in the two models, respectively. Interestingly, UE90 ranks as the second most important variable in GeoGAT-E, but as the second least important in MGWR. This contrast showcases the advantages of GeoGAT-E. For a long time, there has been a paradox between unemployment rates and homicide rates in non-causal analysis: official crime statistics indicate that unemployed individuals and communities with high unemployment rates have higher crime rates; however, cross-sectional studies suggest a negative correlation between unemployment rates and homicide rates (Kapuscinski et al. 1998; South and Cohen 1985). Some studies have pointed out that non-linearity is a potential method to address the homicide rates, though evidence remains weak from their models (McDowall 2002; Raphael and Winter-Ebmer 2001). The strong performance of UE90 in GeoGAT-E demonstrates its capability of effectively separating causal features from shortcut features, providing reliable non-linear interpretations.

5.3 | The Evaluation of Causal Adjustment

Although model evaluation metrics such as R^2 and MAPE can be used to assess the overall model performance, their value in evaluating the quality of the interpretable results obtained through causal adjustment is quite limited. Causal intervention loss measures the mean squared error between the causal features extracted by the model, the shortcut features, and the target label. It is a signal of the model's effective learning of causal features and can be

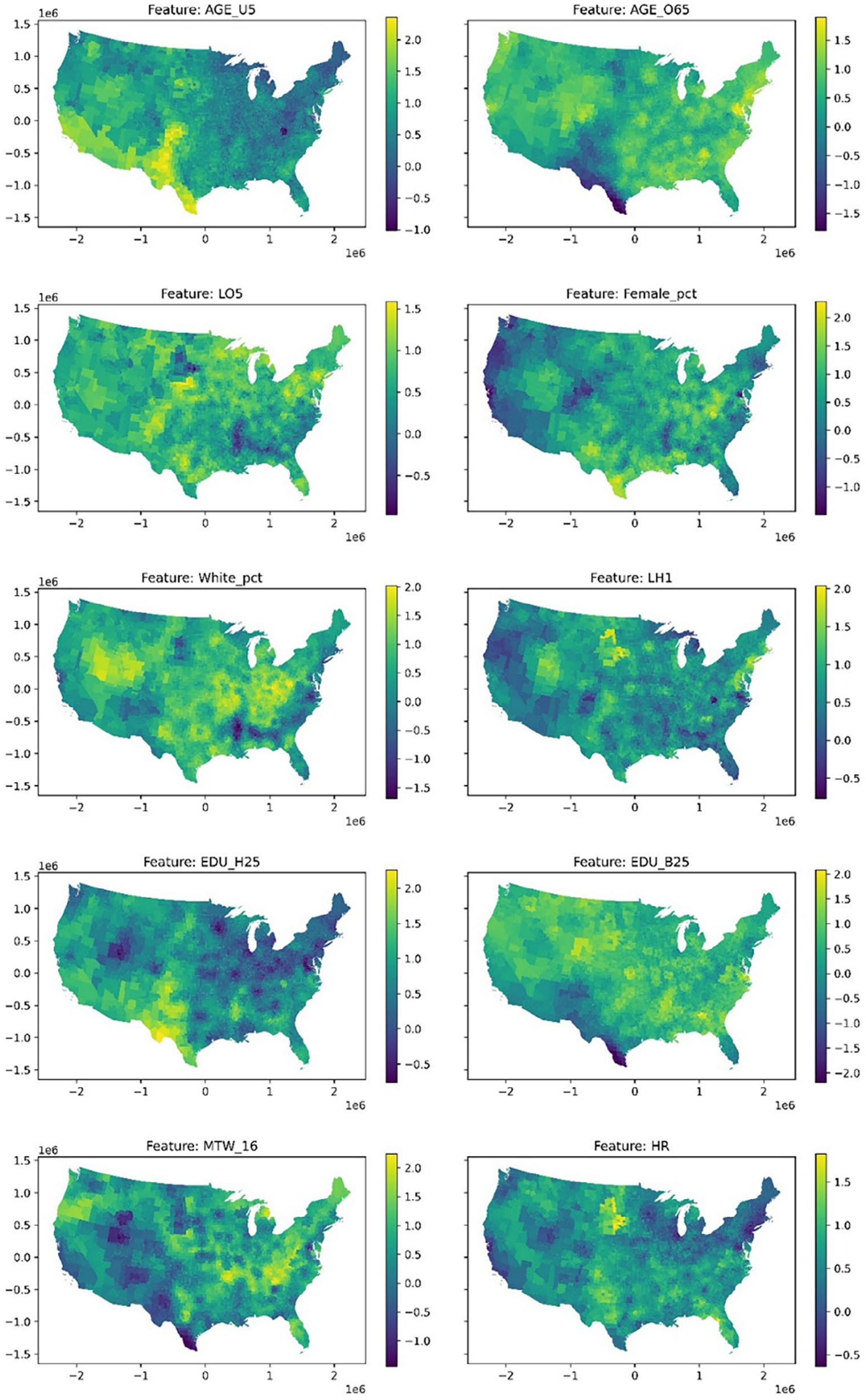

FIGURE 7 | Local feature importance map of Elections.

considered an indicator of the effectiveness of causal feature separation. Evaluating the effectiveness of causal feature separation with causal intervention loss should focus on the downward trend, fluctuation range, and final loss value. As shown in Figure 11, in the Elections dataset, the causal intervention loss is small, with a small fluctuation range and continues to decrease during training, indicating that the model effectively distinguishes and extracts causal and shortcut features. In the Homicides dataset, the causal intervention loss decreases but with a large fluctuates and a relatively large loss value, suggesting that the model may not effectively separate causal and shortcut features. By combining R^2, MAPE, and causal intervention loss, we can conclude that the interpretability of Geo-GAT-E's results is reliable in the Elections dataset, whereas the interpretability of its results in the Homicides dataset may be biased.

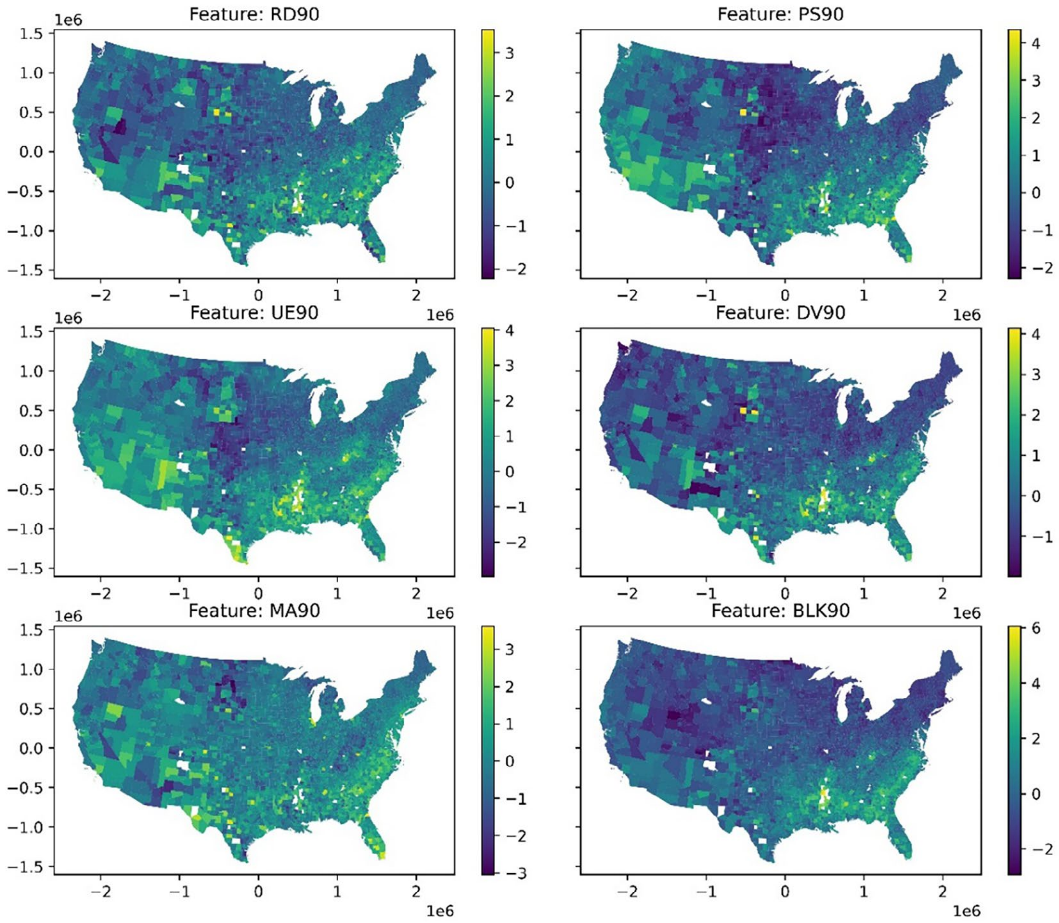

FIGURE 8 | Local feature importance map of Homicides.

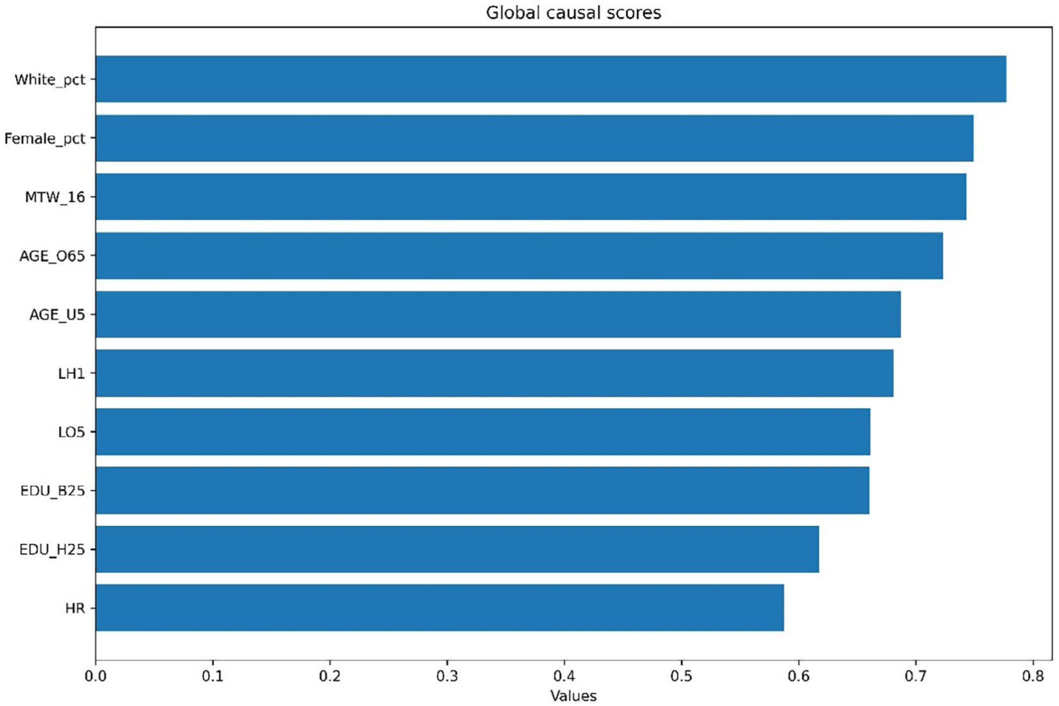

FIGURE 9 | Global feature importance of Elections.

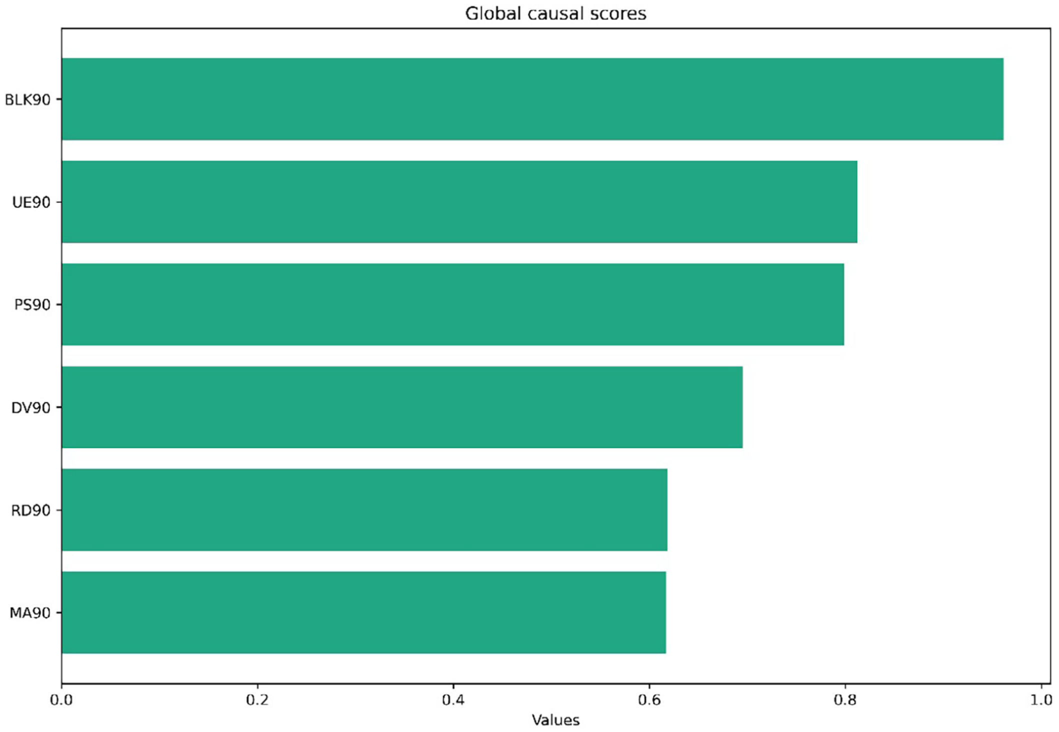

FIGURE 10 | Global feature importance of Homicides.

TABLE 6 | Summary statistics of GeoGAT-E on the Homicides dataset.

	Min	25%	50%	75%	Max	Abs. mean
RD90	−2.477	−0.570	−0.141	0.521	4.662	0.703
PS90	−2.358	−0.668	−0.213	0.529	4.152	0.737
UE90	−2.813	−0.714	−0.203	0.483	4.166	0.766
DV90	−2.176	−0.576	0.019	0.576	4.708	0.722
MA90	−3.405	−0.544	−0.076	0.525	3.770	0.677
BLK90	−2.548	−0.837	0.235	0.567	5.709	0.883

TABLE 7 | Coefficient estimates of MGWR on the Homicides dataset.

	Min	25%	50%	75%	Max	Abs. mean
RD90	−0.144	0.198	0.291	0.391	0.881	0.299
PS90	0.030	0.113	0.131	0.178	0.356	0.151
UE90	−0.083	−0.068	−0.052	0.003	0.069	0.049
DV90	0.027	0.074	0.104	0.131	0.219	0.106
MA90	−0.017	−0.001	0.015	0.021	0.024	0.014
BLK90	0.363	0.387	0.492	0.677	0.746	0.526

6 | Discussion

6.1 | Spatial Similarity for Spatial Analysis

Zhu et al. (2018) systematically summarized and proposed spatial similarity as a promising principle for spatial prediction. Traditionally, spatial similarity has been used to calculate the similarity between known and unknown observations, representing the contribution of known observations to the prediction of unknown ones (Zhu et al. 1997, 2015). While this approach is feasible, it requires modifications to adapt to graph DL models. In this study, inspired by the approach of SRGCNN (Zhu et al. 2022), we

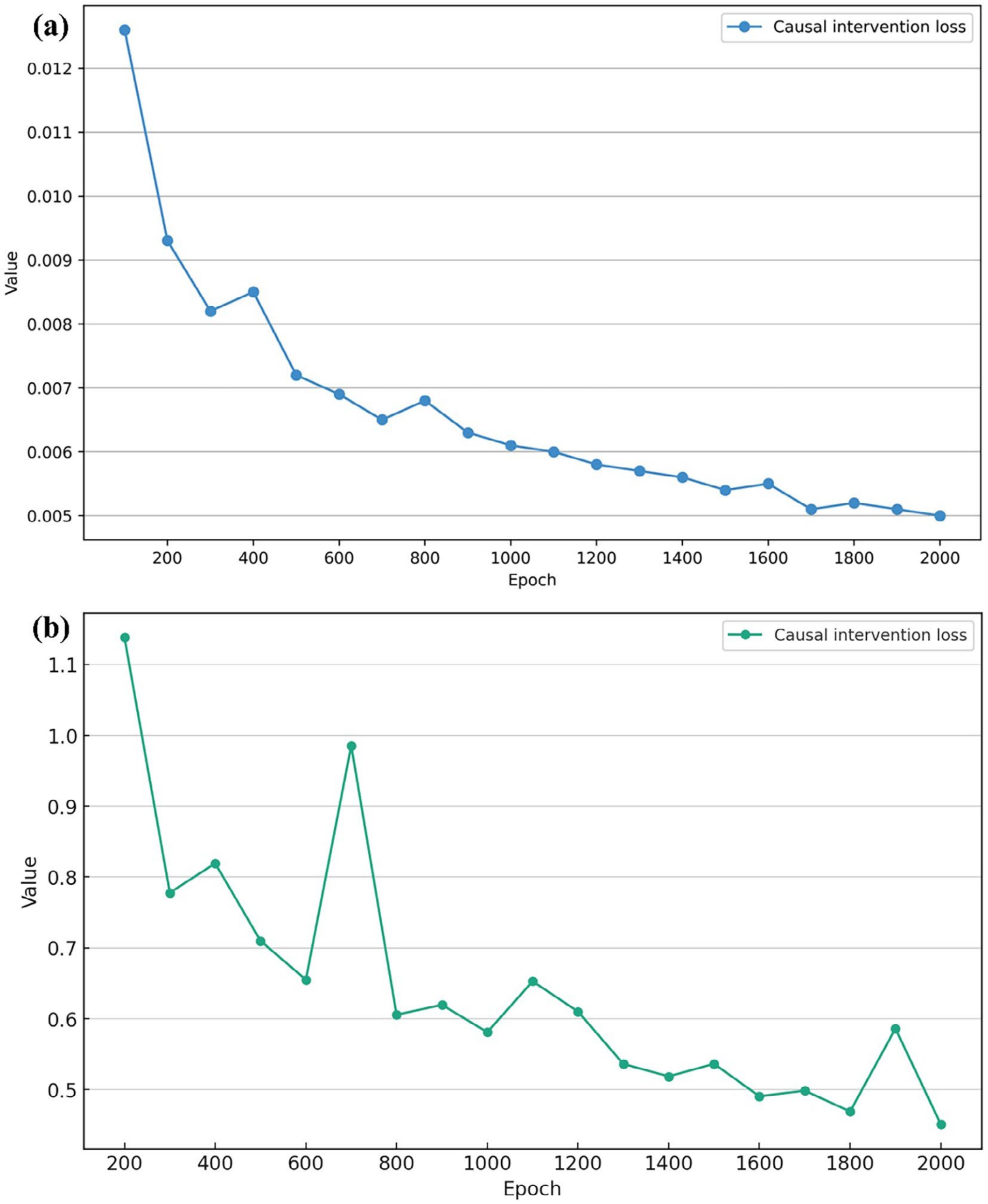

FIGURE 11 | Causal intervention loss over epochs on the datasets: (a) Elections dataset; (b) Homicides dataset.

construct a global subgraph for each observation to meet the inherent requirements of the GAT model based on spatial similarity. For spatial prediction, GeoGAT-P utilizes spatial similarity defined as the combination of the similarity between two nodes themselves and the similarity between their neighborhoods. Although Zhu et al. (2018) mentioned that spatial similarity can consider the similarity between neighborhoods, no practical solution was provided. GeoGAT-P addresses this gap by extending spatial similarity to the GAT model, offering a novel spatial prediction paradigm based on spatial similarity. Additionally, spatial similarity was initially proposed for spatial prediction. GeoGAT-E builds a local model for each observation using the optimal bandwidth, employing spatial similarity as a weighted mechanism instead of spatial distance, which extends spatial similarity to ESDA tasks. According to the performance of GeoGAT-E in Section 5, this approach effectively mitigates the impact of local outliers.

In Sections 4 and 5.1, our study demonstrated that the GeoGAT models, based on spatial similarity, offered significant advantages over the SRGCNN model, which relies on spatial dependence. These advantages are evident in both spatial prediction and ESDA. GeoAI emphasizes the integration of AI within geographical contexts, and recent years have seen a proliferation of research combining AI methodologies with traditional spatial econometric models (Du et al. 2020; Li et al. 2023; Zhu et al. 2022). However, a majority of GeoAI models predominantly focus on incorporating spatial dependence and spatial heterogeneity, potentially conveying to geographers that these are the primary paradigms for the confluence of AI and geography.

In the context of spatial prediction, the assumption of spatial dependence and spatial heterogeneity poses challenges in enhancing the predictive power of GeoAI models across large areas. This is particularly problematic when sufficient samples to measure spatial dependence and spatial heterogeneity are unavailable (Baltagi and Baltagi 2001; Cressie 2015; LeSage and Pace 2014; Zhu et al. 2018). Rather than abandoning the assumption, we advocate for integrating additional geographical theories and methodologies to overcome their limitations. Spatial similarity is more prominent in spatial prediction compared to spatial dependence and spatial heterogeneity. It compensates for the drawbacks of methods based on spatial dependence and spatial heterogeneity, which require large samples and small areas (Zhu et al. 2018; Zhu and Turner 2022). In this study, the excellent spatial prediction capability of GeoGAT-P, even with a small 10% training set size (Table 4), further validates the advantage of spatial similarity in spatial prediction. Spatial dependence, spatial heterogeneity, and spatial similarity are not contradictory; they coexist and complement each other (Zhu et al. 2018; Zhu and Turner 2022). This study combines spatial heterogeneity and spatial similarity to calibrate a local GAT model for ESDA, while also reflecting spatial dependence. GeoGAT-E outperformed SRGCNN, demonstrating that integrating spatial similarity and spatial heterogeneity with other geographical theories is feasible and can enhance the capabilities of spatial analysis against local outliers.

6.2 | Model Interpretability Based on Casual Adjustment

The purpose of interpretability methods in DL is to help users understand and explain the model's decision-making process and the output results, ensuring the model's credibility and transparency (Gilpin et al. 2018; Samek et al. 2017). However, the result explanation in MLP is based on the model's structure and training process, primarily reflecting the influence of features within the neural network rather than their actual causal relationship with the target feature (Guidotti et al. 2018; Lipton 2018). This is because the development of MLP models has largely paralleled the advancement of computer vision (CV), with their design and application being heavily influenced by image processing paradigms (Goodfellow et al. 2016; LeCun et al. 2015; Zhang et al. 2019). In spatial data analysis, feature importance must consider spatial relationships, not merely the weights within the models. Traditional MLP models do not sufficiently account for these aspects, potentially leading to misconceptions about feature importance.

Although some studies have provided insights, for example, Li (2024) proposed GeoShapley, which is based on game theory to capture spatial effects from ML models (Li 2024). The approach requires incorporating spatial locations with strong spatial autocorrelation as variables, which may render the model unreliable (Meyer and Pebesma 2022; Meyer et al. 2018; Meyer et al. 2019). Moreover, this method may exacerbate the effects of spatial confounding, making it difficult to distinguish important relationships (Donegan 2024; Reich et al. 2006). Zhao and Zhang (2024) have indirectly demonstrated that separating causal features from shortcut features using causal theory to enhance a model's interpretability and generalization capability on spatial data is feasible. Besides, many studies have demonstrated the effectiveness of attention weights as a tool for model's interpretability (Vaswani 2017; Lin et al. 2017). In this study, we utilized a causal attention mechanism and backdoor adjustment to distinguish important features from confounding features, reducing the bias caused by noise and increasing the model's credibility. The interpretability then was projected by a linear layer, which is the weights of attention mechanism after causal adjustment. This approach, by using attention weights for interpretation after first applying causal adjustments, theoretically significantly enhances the model's interpretability. It also provided a new insight into interpretability for GeoAI models to understand the complex spatial relationships.

6.3 | Limitations and Future Prospects

We must acknowledge that our models are not perfect, and some limitations need to be addressed in future research. First, although GeoGAT-E can distinguish causal features from confounding features, it is still unable to solve the spatial confounding. Spatial confounding refers to the collinearity between covariates and spatial effects, leading to significant bias in the estimated effects (Clayton et al. 1993; Donegan 2024; Reich et al. 2006). Future work can adopt the framework of Restricted Spatial Regression (Hoffman and Kedron 2023; Reich et al. 2006) to tackle spatial confounding issues in GeoGAT-E. Second, although GeoGAT-E's interpretability has a theoretical foundation and is supported by previous research (Sui et al. 2022; Zhao and Zhang 2024; Vaswani 2017; Wiegreffe and Pinter 2019), the precise accuracy of its interpretability still requires further validation. Due to the lack of theoretical foundations for simulated data, it is challenging for validation. This paper uses causal intervention loss as an indicator to measure the effectiveness of causal feature separation, but this metric may have biases. Future work needs to propose a comprehensive framework to evaluate the interpretable results obtained through causal adjustment. Finally, GeoGAT-E employed two independent graph attention convolutional layers to separately extract causal features and shortcut features. Although this approach is effective in enhancing model interpretability and generalization ability, it has not yet been developed into a causal inference model. The next steps should aim to realize its causal inference capabilities, referencing complex systems theories (Gao, Yang et al. 2023).

7 | Conclusion

This study proposes the GeoGATs for spatial prediction and ESDA, namely GeoGAT-P and GeoGAT-E, respectively. GeoGAT-P, calibrated and extended by spatial similarity, has demonstrated excellent performance in spatial prediction with small samples across large areas, as shown in the U.S. Elections and Homicides datasets. The predictive performance of GeoGAT-P outperforms that of SRGCNN and GNNWR, both of which are calibrated for spatial dependence and spatial heterogeneity, demonstrating that spatial similarity is more suitable to calibrate AI models for small-sample prediction than spatial dependence and spatial heterogeneity. GeoGAT-E is calibrated by spatial heterogeneity and spatial similarity to build the local model and uses the causal attention mechanism to capture and interpret spatial associations, providing a new insight in interpretability and ESDA of GeoAI. GeoGAT-E's performance in the complex U.S. Homicides dataset surpasses SRGCNN and GRF, and significantly outperforms MGWR, evidencing the prospect of spatial similarity for GeoAI in ESDA.

GeoGATs provide scholars in relevant fields with new spatial analysis tools that are capable of handling complex spatial prediction and ESDA tasks. As an endeavor in GeoAI methodology, our work combines the popular GAT model and the crucial spatial similarity principle, offering new insights into the role of spatial similarity in GeoAI models and applications. Future work includes proposing a comprehensive framework to validate and evaluate the interpretable results obtained through causal adjustment, as well as advancing the model toward spatial causal inference.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The codes used in this paper will be shared on GitHub.

Endnotes

[1] Multi-layer perceptron (MLP) is the foundation of most deep learning models, introducing layered architectures, non-linear activations, and backpropagation for training. Modern deep models like CNNs, RNNs, GATs, and Transformers expand on MLP principles to address specialized tasks and data types.

[2] Neighborhood can be identified by Queen contiguity, Rook contiguity, and K-Nearest Neighbors (KNN).

[3] SRGCNN first constructs the spatial adjacency matrix using k-nearest neighbors (KNN) and then applies the graph Laplacian matrix to capture spatial dependence in the model. The code can be found at: https://github.com/dizhu-gis/SRGCNN.

[4] SRGAT constructs the spatial adjacency matrix using KNN, distance or spatial contiguity; and then uses attention mechanism in GAT to capture the spatial dependence.

References

Alshari, E. A., M. B. Abdulkareem, and B. W. Gawali. 2023. "Classification of Land Use/Land Cover Using Artificial Intelligence (ANN-RF)." *Frontiers in Artificial Intelligence* 5: 964279.

Anselin, L. 1995. "Local Indicators of Spatial Association—LISA." *Geographical Analysis* 27, no. 2: 93–115.

Anselin, L. 2002. "Under the Hood Issues in the Specification and Interpretation of Spatial Regression Models." *Agricultural Economics* 27, no. 3: 247–267.

Baldi, P., and P. J. Sadowski. 2013. "Understanding dropout." In *Advances in Neural Information Processing Systems*, vol. 26., 1–9. Curran Associates.

Baltagi, B. H., and B. H. Baltagi. 2001. *A Companion to Theoretical Econometrics*. Wiley Online Library.

Brunsdon, C., A. S. Fotheringham, and M. E. Charlton. 1996. "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity." *Geographical Analysis* 28, no. 4: 281–298.

Brynte, L., J. P. Iglesias, C. Olsson, and F. Kahl. 2024. "Learning Structure-From-Motion With Graph Attention Networks." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4808–4817. Institute of Electrical and Electronics Engineers (IEEE).

Christin, S., E. Hervet, and N. Lecomte. 2019. "Applications for Deep Learning in Ecology." *Methods in Ecology and Evolution* 10, no. 10: 1632–1644.

Clayton, D. G., L. Bernardinelli, and C. Montomoli. 1993. "Spatial Correlation in Ecological Analysis." *International Journal of Epidemiology* 22, no. 6: 1193–1202.

Cortes, C., M. Mohri, and A. Rostamizadeh. 2012. "L2 Regularization for Learning Kernels." ArXiv preprint arXiv:1205.2653.

Cressie, N. 2015. *Statistics for Spatial Data*. John Wiley & Sons.

Donegan, C. 2024. "Plausible Reasoning and Spatial-Statistical Theory: A Critique of Recent Writings on "Spatial Confounding"." *Geographical Analysis* 57, no. 1: 152–172.

Du, Z., Z. Wang, S. Wu, F. Zhang, and R. Liu. 2020. "Geographically Neural Network Weighted Regression for the Accurate Estimation of Spatial Non-stationarity." *International Journal of Geographical Information Science* 34, no. 7: 1353–1377.

Fotheringham, A. S., H. Yu, L. J. Wolf, T. M. Oshan, and Z. Li. 2022. "On the Notion of 'Bandwidth' in Geographically Weighted Regression Models of Spatially Varying Processes." *International Journal of Geographical Information Science* 36, no. 8: 1485–1502.

Gao, B., J. Yang, Z. Chen, et al. 2023. "Causal Inference From Cross-Sectional Earth System Data With Geographical Convergent Cross Mapping." *Nature Communications* 14, no. 1: 5875. https://doi.org/10.1038/s41467-023-41619-6.

Gao, S., Y. Hu, and W. Li. 2023. *Handbook of Geospatial Artificial Intelligence*. CRC Press.

Georganos, S., T. Grippa, A. Niang Gadiaga, et al. 2021. "Geographical Random Forests: A Spatial Extension of the Random Forest Algorithm to Address Spatial Heterogeneity in Remote Sensing and Population Modelling." *Geocarto International* 36, no. 2: 121–136. https://doi.org/10.1080/10106049.2019.1595177.

Gilpin, L. H., D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. 2018. "Explaining Explanations: An Overview of Interpretability of Machine Learning." In *2018 Ieee 5th International Conference on Data Science and Advanced Analytics (Dsaa)*, 80–89. IEEE.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. MIT Press.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. 2018. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys (CSUR)* 51, no. 5: 1–42.

Harris, P., A. Fotheringham, R. Crespo, and M. Charlton. 2010. "The Use of Geographically Weighted Regression for Spatial Prediction: An Evaluation of Models Using Simulated Data Sets." *Mathematical Geosciences* 42: 657–680.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hoffman, T. D., and P. Kedron. 2023. "Controlling for Spatial Confounding and Spatial Interference in Causal Inference: Modelling Insights From a Computational Experiment." *Annals of GIS* 29, no. 4: 517–527.

Huber, F. 2018. *A Logical Introduction to Probability and Induction*. Oxford University Press.

Hulland, J. 1999. "Use of Partial Least Squares (PLS) in Strategic Management Research: A Review of Four Recent Studies." *Strategic Management Journal* 20, no. 2: 195–204.

Janowicz, K., S. Gao, G. McKenzie, Y. Hu, and B. Bhaduri. 2020. *GeoAI: Spatially Explicit Artificial Intelligence Techniques for Geographic Knowledge Discovery and Beyond*. Taylor & Francis.

Kapuscinski, C. A., J. Braithwaite, and B. Chapman. 1998. "Unemployment and Crime: Toward Resolving the Paradox." *Journal of Quantitative Criminology* 14: 215–243.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *Nature* 521, no. 7553: 436–444.

LeSage, J. P., and R. K. Pace. 2014. “The Biggest Myth in Spatial Econometrics.” *Econometrics* 2, no. 4: 217–249.

Li, K., J. Zhu, A. R. Ives, V. C. Radeloff, and F. Wang. 2023. “Semiparametric regression for spatial data via deep learning.” *Spatial Statistics* 57: 100777.

Li, Z. 2024. “Geoshapley: A Game Theory Approach to Measuring Spatial Effects in Machine Learning Models.” *Annals of the American Association of Geographers* 114, no. 7: 1365–1385.

Li, Z., A. S. Fotheringham, T. M. Oshan, and L. J. Wolf. 2020. “Measuring Bandwidth Uncertainty in Multiscale Geographically Weighted Regression Using Akaike Weights.” *Annals of the American Association of Geographers* 110, no. 5: 1500–1520.

Lin, Z., M. Feng, C. N. D. Santos, et al. 2017. “A Structured Self-Attentive Sentence Embedding.” ArXiv preprint arXiv:1703.03130.

Lipton, Z. C. 2018. “The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery.” *Queue* 16, no. 3: 31–57.

Liu, P., and F. Biljecki. 2022. “A Review of Spatially-Explicit Geoai Applications in Urban Geography.” *International Journal of Applied Earth Observation and Geoinformation* 112: 102936.

Louizos, C., U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. 2017. “Causal Effect Inference With Deep Latent-Variable Models.” *Advances in Neural Information Processing Systems* 30. https://proceedings.neurips.cc/paper/2017/file/94b5bde6de888ddf9cde6748ad2523d1-Paper.pdf.

McDowall, D. 2002. “Tests of Nonlinear Dynamics in US Homicide Time Series, and Their Implications.” *Criminology* 40, no. 3: 711–736.

Meyer, H., and E. Pebesma. 2022. “Machine Learning-Based Global Maps of Ecological Variables and the Challenge of Assessing Them.” *Nature Communications* 13, no. 1: 2208.

Meyer, H., C. Reudenbach, T. Hengl, M. Katurji, and T. Nauss. 2018. “Improving Performance of Spatio-Temporal Machine Learning Models Using Forward Feature Selection and Targetoriented Validation.” *Environmental Modelling & Software* 101: 1–9.

Meyer, H., C. Reudenbach, S. Wollauer, and T. Nauss. 2019. “Importance of Spatial Predictor Variable Selection in Machine Learning Applications–Moving From Data Reproduction to Spatial Prediction.” *Ecological Modelling* 411: 108815.

Nikparvar, B., and J. C. Thill. 2021. “Machine Learning of Spatial Data.” *ISPRS International Journal of Geo-Information* 10, no. 9: 600.

Oshan, T. M., Z. Li, W. Kang, L. J. Wolf, and A. S. Fotheringham. 2019. “Mgwr: A Python Implementation of Multiscale Geographically Weighted Regression for Investigating Process Spatial Heterogeneity and Scale.” *ISPRS International Journal of Geo-Information* 8, no. 6: 269.

Özgür, A., and F. Nar. 2020. “Effect of Dropout Layer on Classical Regression Problems.” In *2020 28th Signal Processing and Communications Applications Conference (SIU)*, 1–4. IEEE.

Ploton, P., F. Mortier, M. Réjou-Méchain, et al. 2020. “Spatial Validation Reveals Poor Predictive Performance of Large-Scale Ecological Mapping Models.” *Nature Communications* 11, no. 1: 4540. https://doi.org/10.1038/s41467-020-18321-y.

Raphael, S., and R. Winter-Ebmer. 2001. “Identifying the Effect of Unemployment on Crime.” *Journal of Law and Economics* 44, no. 1: 259–283.

Reich, B. J., J. S. Hodges, and V. Zadnik. 2006. “Effects of Residual Smoothing on the Posterior of the Fixed Effects in Disease-Mapping Models.” *Biometrics* 62, no. 4: 1197–1206.

Samek, W., T. Wiegand, and K.-R. Mu¨ller. 2017. “Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models.” ArXiv preprint arXiv:1708.08296.

South, S. J., and L. E. Cohen. 1985. “Unemployment and the Homicide Rate: A Paradox Resolved?” *Social Indicators Research* 17: 325–343.

Sui, Y., X. Wang, J. Wu, M. Lin, X. He, and T.-S. Chua. 2022. “Causal Attention for Interpretable and Generalizable Graph Classification.” In *Proceedings of the 28th Acm Sigkdd Conference on Knowledge Discovery and Data Mining*, 1696–1705. Association for Computing Machinery (ACM).

Vaswani, A. 2017. “Attention is all you need.” ArXiv preprint arXiv:1706.03762.

Velickovic, P., G. Cucurull, A. Casanova, et al. 2017. “Graph Attention Networks.” *Stat* 1050, no. 20: 10–48550.

Vrahatis, A. G., K. Lazaros, and S. Kotsiantis. 2024. “Graph Attention Networks: A Comprehensive Review of Methods and Applications.” *Future Internet* 16, no. 9: 318.

Ward, M. D., and K. S. Gleditsch. 2018. *Spatial regression models*. Sage Publications.

Wiegreffe, S., and Y. Pinter. 2019. “Attention Is Not Not Explanation.” ArXiv preprint arXiv:1908.04626.

Wooldridge, M. 2021. *A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going*. Flatiron Books.

Xie, S., L. Li, and Y. Zhu. 2024. “Anomaly Detection for Multivariate Time Series in IoT Using Discrete Wavelet Decomposition and Dual Graph Attention Networks.” *Computers & Security* 146: 104075.

Xu, Y., and R. Zuo. 2024. “An Interpretable Graph Attention Network for Mineral Prospectivity Mapping.” *Mathematical Geosciences* 56, no. 2: 169–190.

Yuan, Q., H. Shen, T. Li, et al. 2020. “Deep Learning in Environmental Remote Sensing: Achievements and Challenges.” *Remote Sensing of Environment* 241: 111716. https://doi.org/10.1016/j.rse.2020.111716.

Zhang, C., S. Bengio, M. Hardt, B. Recht, and O. Vinyals. 2021. “Understanding Deep Learning (Still) Requires Rethinking Generalization.” *Communications of the ACM* 64, no. 3: 107–115.

Zhang, W., Y. Yu, Y. Qi, F. Shu, and Y. Wang. 2019. “Short-Term Traffic Flow Prediction Based on Spatio-Temporal Analysis and Cnn Deep Learning.” *Transportmetrica A: Transport Science* 15, no. 2: 1688–1711.

Zhao, K., and L. Zhang. 2024. “Causality-Inspired Spatial-Temporal Explanations for Dynamic Graph Neural Networks.” In *The Twelfth International Conference on Learning Representations*. Published on OpenReview.

Zhao, P., Z. Li, Z. Xiao, S. Jiang, Z. He, and M. Zhang. 2023. “Spatiotemporal Characteristics and Driving Factors of CO2 Emissions From Road Freight Transportation.” *Transportation Research Part D: Transport and Environment* 125: 103983.

Zheng, L., and W. Lu. 2024. “Urban Micro-Scale Street Thermal Comfort Prediction Using a ‘Graph Attention network’ Model.” *Building and Environment* 262: 111780.

Zhu, A., J. Liu, F. Du, et al. 2015. “Predictive Soil Mapping With Limited Sample Data.” *European Journal of Soil Science* 66, no. 3: 535–547. https://doi.org/10.1111/ejss.12244.

Zhu, A. X., L. Band, R. Vertessy, and B. Dutton. 1997. “Derivation of Soil Properties Using a Soil Land Inference Model (SoLIM).” *Soil Science Society of America Journal* 61, no. 2: 523–533.

Zhu, A.-X., G. Lu, J. Liu, C.-Z. Qin, and C. Zhou. 2018. “Spatial Prediction Based on Third Law of Geography.” *Annals of GIS* 24, no. 4: 225–240.

Zhu, A.-X., and M. Turner. 2022. “How Is the Third Law of Geography Different?” *Annals of GIS* 28, no. 1: 57–67.

Zhu, D., Y. Liu, X. Yao, and M. M. Fischer. 2022. “Spatial Regression Graph Convolutional Neural Networks: A Deep Learning Paradigm for Spatial Multivariate Distributions.” *GeoInformatica* 26, no. 4: 645–676.

Appendix A

The Sensitivity of the Weight of Neighborhood in GeoGAT-P

In Figure A1, we can find that the optimal weight of neighborhood is 1.

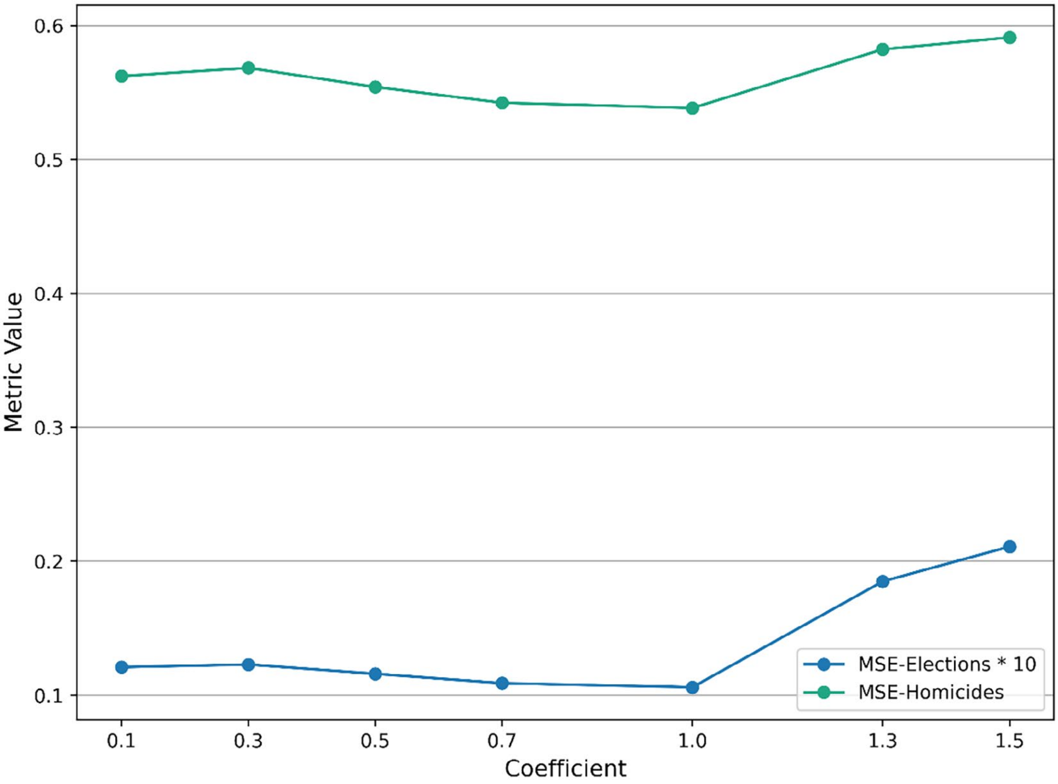

FIGURE A1 | The performances of GeoGAT-P across different weights on neighborhood.

Appendix B

The Selection of the Number of Similar Observations in GeoGAT-P

In Figure B1, we observe that selecting the 10 most similar observations as the neighborhood for GeoGAT-P yields the best predictive performance across both datasets.

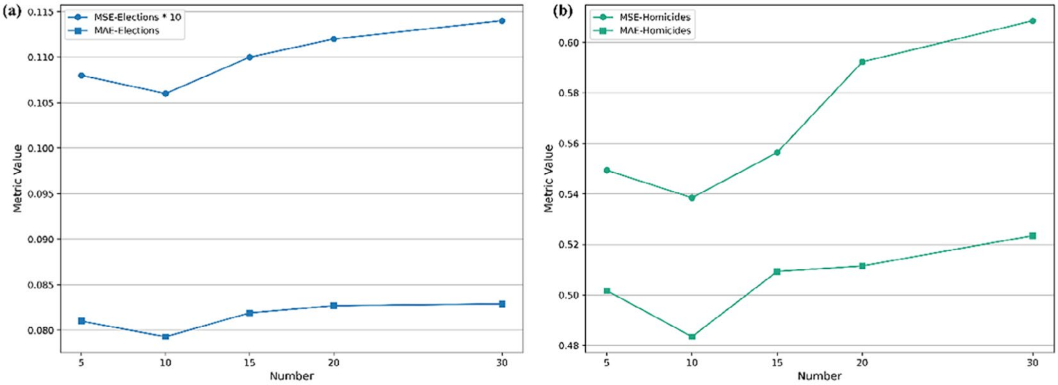

FIGURE B1 | The performances of GeoGAT-P across two datasets as the number of neighbors changes. (a) Elections dataset; (b) Homicides dataset.

Appendix C

GeoGATs′ Computational Performances on Spatial Prediction and ESDA

GeoGAT-P for Spatial Prediction

The computational performance of the models was assessed in terms of runtime and RAM usage. GNNWR emerged as the most efficient, achieving the fastest runtime and minimal RAM consumption. GeoGAT-P followed closely, delivering performance nearly on par with GNNWR. On the other hand, SRGCNN demonstrated the least favorable computational performance among all models.

TABLE C1 | Computational cost in spatial prediction across all models.

	Elections		Homicides	
Model	**Time (min)**	**RAM consumption (GB)**	**Time (min)**	**RAM consumption (GB)**
GNNWR	5	1.6	5	1.5
SRGCNN	10	2.5	10	2.9
GeoGAT-P	7	1.9	7	1.8

GeoGAT-E for Exploratory Spatial Data Analysis

In the ESDA task, GeoGAT-E achieved the best computational efficiency, with the shortest runtime and minimal RAM usage, followed by SRGCNN. However, GRF and MGWR exhibited extremely high computational costs, posing challenges for processing large datasets.

TABLE C2 | Computational cost in ESDA across all models.

	Elections		Homicides	
Model	**Time**	**RAM consumption (GB)**	**Time**	**RAM consumption (GB)**
GRF	3 h	25	3 h	27
SRGCNN	15 min	3.1	14 min	3.2
MGWR	1.6 h	4.2	2 h	4.8
GeoGAT-P	10 min	2.5	10 min	2.6