OpenCarbon: A Contrastive Learning-based Cross-Modality Neural Approach for High-Resolution Carbon Emission Prediction Using Open Data

Jinwei Zeng[1], **Yu Liu**[1], **Guozhen Zhang**[2], **Jingtao Ding**[1], **Yuming Lin**[1], **Jian Yuan**[1], **Yong Li**[*1]
[1]Tsinghua University
[2]TsingRoc
zengjw17@gmail.com, liyong07@tsinghua.edu.cn

Abstract

Accurately estimating high-resolution carbon emissions is crucial for effective emission governance and mitigation planning. While conventional methods for precise carbon accounting are hindered by substantial data collection efforts, the rise of open data and advanced learning techniques offers a promising solution. Once an open data-based prediction model is developed and trained, it can easily infer emissions for new areas based on available open data. To address this, we incorporate two modalities of open data, satellite images and point-of-interest (POI) data, to predict high-resolution urban carbon emissions, with satellite images providing macroscopic and static and POI data offering fine-grained and relatively dynamic functionality information. However, estimating high-resolution carbon emissions presents two significant challenges: the intertwined and implicit effects of various functionalities on carbon emissions, and the complex spatial contiguity correlations that give rise to the agglomeration effect. Our model, OpenCarbon, features two major designs that target the challenges: a cross-modality information extraction and fusion module to extract complementary functionality information from two modules and model their interactions, and a neighborhood-informed aggregation module to capture the spatial contiguity correlations. Extensive experiments demonstrate our model's superiority, with a significant performance gain of 26.6% on R^2. Further generalizability tests and case studies also show OpenCarbon's capacity to capture the intrinsic relation between urban functionalities and carbon emissions, validating its potential to empower efficient carbon governance and targeted carbon mitigation planning. Codes and data are available: https://github.com/JinweiZzz/OpenCarbon.

1 Introduction

Accurately accounting for high-resolution urban carbon emissions has become an increasingly critical issue [Stechemesser and Guenther, 2012]. Urban carbon emissions now account for more than 70% of global emissions [Agreement, 2015; Bank, 2021] and continue to rise. To fully understand the sources of these emissions and effectively design carbon-reduction policies targeting high-emission areas, it is crucial to account for urban carbon emissions at a high resolution [Stechemesser and Guenther, 2012].

Despite the importance of emission accounting, traditional carbon accounting methods face a trade-off between accuracy and the effort required for data acquisition [Cai *et al.*, 2018; Olivier *et al.*, 1999; Dai *et al.*, 2016]. Existing approaches to carbon emission accounting generally fall into two categories: bottom-up and top-down [Hutchins *et al.*, 2017; Böhringer and Rutherford, 2008; Böhringer, 1998]. The bottom-up approach, which relies on point emission statistics from sensors or fine-grained activity data [Gurney *et al.*, 2020; Gurney *et al.*, 2009], is accurate but requires extensive data collection, making it unsustainable. In contrast, the top-down approach distributes regional sectoral emission totals—calculated from fuel consumption statistics—across high-resolution areas using proxy factors [Olivier *et al.*, 1994]. However, the ideal assumption of positive correlations between proxies and emissions often leads to inaccuracies. Given the strengths of machine learning and deep learning in identifying complex correlations, along with the vast availability of open data—which is widely used in socioeconomic prediction tasks [Yeh *et al.*, 2020; Aiken *et al.*, 2022; Xu *et al.*, 2020; Wang *et al.*, 2016]—developing a model that predicts carbon emissions using open data presents a promising solution. Once trained, the model can accurately infer a region's emissions using readily available open data, maintaining accuracy while significantly reducing costly data collection efforts.

Nevertheless, directly applying existing open data-based socioeconomic prediction methods to predict carbon emissions fails to capture the unique and challenging characteristics of carbon emission spatial distribution: the functional effect and spatial agglomeration effect. While existing socioeconomic indicators typically reflect one type of activity or urban functionality, carbon emissions result from the combination of diverse functionalities that have distinct effects on emissions [Dhakal, 2009; Liu *et al.*, 2020] (Fig 1(a)). For instance, traffic roads generate transportation carbon emissions as vehicles burn petrol [Wang *et al.*, 2015], while residential areas involve heating and cooking activities that consume

[*]Corresponding author.

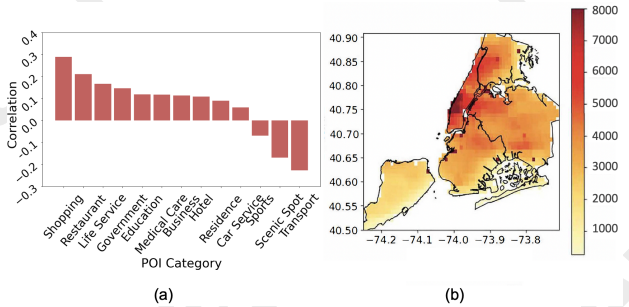

Figure 1: Illustrations of the (a)functional effect and (b)spatial agglomeration effect of urban carbon emission distribution. (a) Correlation between the shares of different POI types and carbon emissions in Beijing. (b) Spatial distribution of New York's carbon emissions.

natural gas and liquefied petroleum gas [Nejat *et al.*, 2015; Zhang *et al.*, 2015]. Learning the implicit and coupled relationships between such functionalities and carbon emissions is complex and challenging. Meanwhile, urban carbon emissions exhibit a strong and unique spatial agglomeration effect, originating from the continuity in city functionality layouts [Han *et al.*, 2018; Wang *et al.*, 2018; Wang *et al.*, 2020]. As shown in Figure 1(b), adjacent areas in New York City have similar emission levels, and the overall carbon emission distribution in the city exhibits continuity. Characterizing such continuity is challenging but essential.

To model the unique characteristics of carbon emissions, we incorporate two types of open data—satellite images and point-of-interest (POI) data—to predict high-resolution spatial carbon emissions for two main reasons. First, satellite images and POI data provide complementary insights into a region's functionalities. Satellite images offer a relatively macroscopic yet comprehensive overview of how the city's static functional zones are distributed [Xu *et al.*, 2020; Zhao *et al.*, 2022], while POI data captures the fine-grained dynamic intensity of human activities, reflecting the rapid and continuous updates of facilities within urban functional zones [Fan *et al.*, 2018]. Second, while satellite images convey information about the spatial layout of each functional zone, POI data is typically geo-tagged and contains valuable location-specific information. Together, these data types form the foundation for modeling the spatial correlations between urban functionalities and activities. Therefore, our primary goal is to effectively leverage these two data types to model the unique and challenging aspects of estimating high-resolution urban carbon emissions.

To bridge the gap, we propose our model, **OpenCarbon**, to predict high-resolution urban carbon emissions with open data of satellite images and POI data. To model the coupled and implicit functional effects from the complementary open data modalities, we propose a cross-modality information extraction and fusion module. This module first extracts information from each modality, explicitly modeling the distinct effects of different functionalities using a function-dimension-wise attention mechanism. It then extracts complementary functional information across the two modalities through a contrastive loss design, aiming to enable each modality to borrow underrepresented information from the other. To capture the spatial contiguity of functionality layouts underlying the spatial agglomeration effect, OpenCarbon features a neighborhood-informed agglomeration modeling module that uses convolutional layers to extract the global context of the neighborhood and a cross-attention mechanism to model grid-neighborhood interactions. Our contribution may be summarized into three parts:

- We tackle the critical challenge of data collection in high-accuracy, high-resolution carbon accounting by developing an open data-based model that provides accurate carbon estimation while significantly reducing the data collection burden.
- Recognizing the unique functional and spatial agglomeration effects of urban carbon emissions, we propose a robust neighborhood-informed attentive neural network with contrastive learning for cross-modality fusion that achieves complementary modality extraction and spatial contiguity modeling.
- Extensive experiments on three large-scale real-world datasets validate our model's superiority over existing state-of-the-art methods, with an average increase of 26.6% in terms of R^2, showing our model's potential to facilitate convenient high-resolution carbon estimation and the carbon governance that follows.

2 Related Works

2.1 Conventional High-resolution Carbon Emission Accounting

High-resolution accounting of carbon emissions is essential for effective, targeted carbon emission governance. Traditional carbon emission accounting methods are generally classified into two categories: bottom-up and top-down approaches [Böhringer and Rutherford, 2008]. Bottom-up methods aggregate monitored point emission data to the desired resolution [Gurney *et al.*, 2020; Gurney *et al.*, 2009], but they require extensive data collection and are not easily scalable across large areas. In contrast, top-down methods distribute regional carbon emissions, typically calculated from regional fuel consumption statistics, to the target resolution using proxy factors [Olivier *et al.*, 1994], such as population. However, because carbon emissions are influenced by a range of factors, relying solely on proxy factors in top-down approaches often leads to inaccuracies. Therefore, we conclude that conventional high-resolution carbon accounting methods face a dilemma between accuracy and the substantial effort required for data collection.

2.2 Carbon Emission Prediction with Open Data

With the development of machine learning and deep learning and the boom in open data, researchers made a large trial in incorporating them to relieve the data collection pressure [Yang *et al.*, 2020; Lu *et al.*, 2017; Chen *et al.*, 2024; Wu *et al.*, 2022]. Yang et al. [Yang *et al.*, 2020] ensembled several multi-layer perception models to predict city-level emissions with nighttime light satellite imagery. Chen et al. [Chen *et al.*, 2024] constructed a graph neural network

to model the spatiotemporal dependencies between nearby counties or districts, predicting temporally fine-grained emissions using available monthly emissions statistics. However, most of these methods target region-level or city-level carbon emissions and may not be suitable for high-resolution prediction due to the limited resolution of the data they involve, or because they inadequately characterize fine-grained functional and spatial interaction information. Therefore, we conclude that the potential of high-resolution open data for predicting carbon emissions remains underdeveloped.

2.3 Socioeconomic Prediction with Satellite Images and POI Distribution Data

Since our solution uses satellite images and POI distribution data as inputs, we review existing works that leverage these data types. Most studies incorporating POI data represent POIs as a distribution count vector. Wang et al.[Wang *et al.*, 2016] developed a neural network model to predict neighborhood crime rates using the POI vector, while Yu et al.[Yu *et al.*, 2018] used POI counts to predict smartphone application usage. For satellite image-related works, visual encoder networks have been applied to learn task-specific image representations [Yeh *et al.*, 2020; Perez *et al.*, 2017]. Researchers have also used unsupervised and self-supervised methods to create unified representations for various socioeconomic tasks. Jean et al.[Jean *et al.*, 2019] designed a triplet loss function to enhance representation similarity for nearby grids, and Xi et al.[Xi *et al.*, 2022] proposed similarity metrics for geo-adjacent grids and grids with similar POI distributions to improve grid semantics. However, to the best of our knowledge, no work has specifically explored the modeling potential of these two complementary data sources for high-resolution carbon emissions. The challenge of effectively extracting complementary information and modeling the implicit functional and spatial agglomeration effects remains underexplored.

3 Preliminaries & Problem Statement

3.1 Definitions

As our work incorporates two types of open data, the satellite image, and POI distribution data, we provide their definition and collection sources here.

Satellite Image: Satellite images are overhead-view pictures of the Earth's surface taken by satellites orbiting the planet. The recent development of remote sensing has enabled the acquisition of **global, high-resolution, and safe** satellite images [Campbell and Wynne, 2011]. Our data source is ArcGIS[1], whose maximum pixel resolution is 1.19m.

POI Distribution: Point-of-interest(POI) data is the geo-tagged facility distribution data that is widely available in various map services and data providers. Some crowd-sourcing map service providers, such as OpenStreetMap, provide access to the global POI data [Zhou *et al.*, 2022]. In this study, we use public POI data with location and category information provided from SafeGraph[2] and Tencent Map.

[1]https://www.arcgis.com/home

[2]https://docs.safegraph.com/docs/places

Notably, all two data modalities are sourced from publicly available, privacy-protected sources with no security risks.

3.2 Problem Statement

We target urban carbon emissions at a spatial granularity that has been paid much attention to in carbon emission governance, the 1km $\times$ 1km grids. For each grid i in the grid set $\mathcal{A}$, given its corresponding satellite image and POI distribution data, our task is to predict its yearly carbon emission.

4 Method: OpenCarbon

In this section, we present our method, OpenCarbon (Figure 2), which predicts high-resolution carbon emissions using open data. OpenCarbon comprises two main components: (1) a cross-modality information extraction and fusion module that leverages complementary functional information from two modalities and models their interactions, and (2) a neighborhood-informed aggregation module designed to capture the continuity of city functionality layouts, thereby modeling the agglomeration effect. In the following sections, we provide a detailed introduction to these two modules and the training process.

4.1 Cross-modality Information Extraction and Fusion

Single-modality Representation Learning

Satellite images offer a broad overview of the functional layouts and their scale, whereas POI data provides complementary, detailed insights into the functionality and activity intensity within each layout. For example, while satellite images provide an immediate sense of a shopping mall's scale, they offer little insight into the composition of its internal store categories. POI data supplements this understanding. Therefore, aligning the observation scales of both modalities is crucial, allowing our model to effectively match corresponding information and develop a comprehensive representation of the area's functionalities.

We use satellite image pixels as the fundamental unit of granularity to organize the geo-tagged POI data into a three-dimensional matrix with consistent pixel dimensions. The first two dimensions represent the spatial distribution of the POIs, while the third encodes their types. The matrix is constructed by mapping each POI to its corresponding matrix element based on its position within the grid. In this way, we obtain the grid-level matrix inputs of the satellite images and POI distribution.

For the satellite image modality, given its visual nature, we incorporate the ResNet-18 [He *et al.*, 2016] architecture to obtain the embedding. Since the POI matrix is relatively sparse, we apply multiple layers of CNNs to obtain a representation of the POI distribution. However, since carbon emissions result from the aggregation of various activities, with distinct and implicit relationships between each activity type and its corresponding emission, we introduce a squeeze-and-excitation (SE) block [Hu *et al.*, 2018] into every CNN layer to account for functionality heterogeneity in the POI distribution representation.

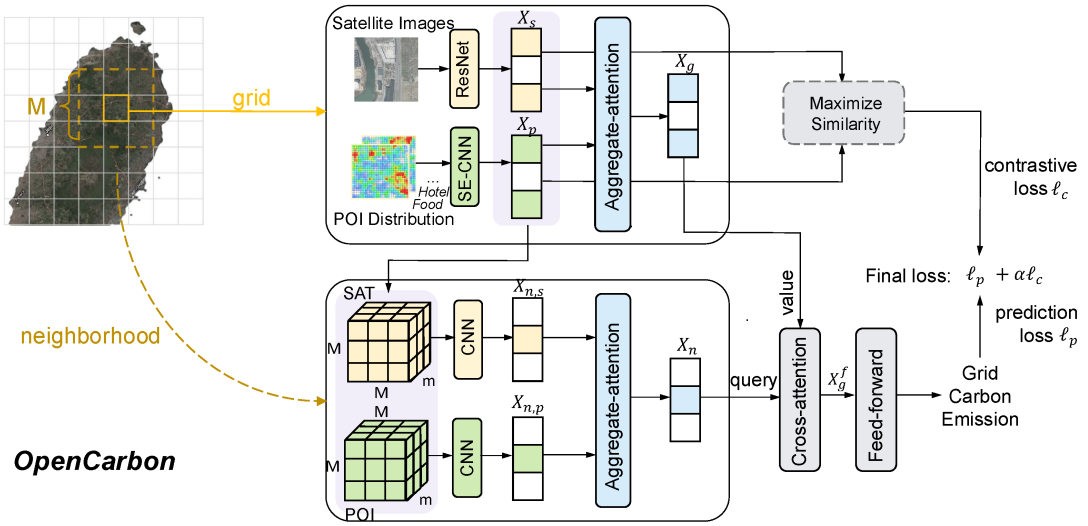

Figure 2: The overall framework of our OpenCarbon model.

Specifically, for any input $X \in \mathbb{R}^{H\times W\times C}$ to the CNN layer, a global average pooling layer serves as the squeeze function to capture the global information of each channel, which can be formulated as

$$z_c = \frac{1}{H \times W}(\sum_h \sum_w X_{hwc}). \tag{1}$$

An excitation function is then introduced to generate attention weights for different channels, scaling the input along the function dimension. The process can be expressed as follows:

$$s_c = \text{Sigmoid}(W_2 \cdot \text{ReLU}(W_1 z_c)), \tag{2}$$

$$X'\cdot,\cdot,c = s_c \cdot X\cdot,\cdot,c, \quad c = 1, 2, ..., C. \tag{3}$$

Here, $W_1 \in \mathbb{R}^{\frac{C}{n}\times C}$ and $W_2 \in \mathbb{R}^{C\times\frac{C}{n}}$ are the learnable projection matrices, and n represents the tunable compression ratio.

Cross-modality Complementary Information Fusion

Once each modality has been represented individually, the next step is to enable each modality to learn complementary information from the other and effectively fuse the two. To accomplish this, we employ contrastive learning, aiming to allow each modality to borrow and encode missing or under-represented information from the other. Specifically, denoting the input from the satellite image modality as X_s and that from the POI modality as X_p, we introduce the well-acknowledged contrastive learning loss, NT-Xent (Normalized Temperature-scaled Cross Entropy) loss [Chen *et al.*, 2020], which can be formulated as

$$\ell_{c,i} = -\log \frac{\exp\left(\frac{\text{sim}(\mathbf{X}_{s,i},\mathbf{X}_{p,i})}{\tau}\right)}{\sum_{k=1}^{N} 1_{[k\neq i]} \exp\left(\frac{\text{sim}(\mathbf{X}_{s,i},\mathbf{X}_{p,k})}{\tau}\right)} - \log \frac{\exp\left(\frac{\text{sim}(\mathbf{X}_{s,i},\mathbf{X}_{p,i})}{\tau}\right)}{\sum_{k=1}^{N} 1_{[k\neq i]} \exp\left(\frac{\text{sim}(\mathbf{X}_{s,k},\mathbf{X}_{p,i})}{\tau}\right)}.$$

Here $\text{sim}(\mathbf{x}_i, \mathbf{x}_j)$ represents the cosine similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. τ is the temperature parameter, used to adjust the distribution of similarity values. N represents the number of samples in a batch.

After obtaining the complementary representations, we apply the aggregate-attention mechanism to adjust the weights of the two modalities during the representation fusion. Specifically, the mechanism maps the representations to a score space and adaptively learns the scores so that the more informative modality receives a higher weight. The attention weights are computed as:

$$\begin{aligned} m_k &= a \cdot \text{Tanh}(W \cdot X_k + b), k \in \{s, p\}, \\ s_k &= \frac{\exp(m_k)}{\sum_{k\in\{s,i\}} \exp(m_k)}, k \in \{s, i\}. \end{aligned} \tag{4}$$

Here W is a learnable matrix, and a and b are learnable vectors. The fused grid-level representation X_g is

$$X_g = \sum_{k\in\{s,i\}} s_k \cdot X_k \tag{5}$$

4.2 Neighborhood-informed Agglomeration Modeling

While a graph structure can represent grid adjacency, it does not capture the relative position information between each grid, as the neighbors of a graph node are unordered. Therefore,

we turn to convolutional layers to model the global representation of the neighborhood. Specifically, in single-modality representation learning, we obtain grid-level representations for satellite images $X_s \in \mathbb{R}^m$ and POI data $X_p \in \mathbb{R}^m$. Considering the $M \times M$ grids surrounding the target grid as the contextual range (where M is a hyperparameter), we concatenate the representations of the two modalities into two matrices, $X_{n,s}$ and $X_{n,p}$, each of size $\mathbb{R}^{M \times M \times m}$. Each matrix is passed through several convolutional layers to learn the spatial agglomeration correlations of its neighborhood. The global representations are then aggregated using the previously introduced aggregate-attention mechanism. This process results in a neighborhood-level representation X_n of the grid.

To capture the spatial continuity underlying the spatial agglomeration effect, we introduce the standard cross-attention mechanism. We use the contextual neighborhood representation as the query vector to model the spatial correlations and interactions. In this way, we obtain the final grid representation X_g^f, which can be formulated as:

$$X_g^f = \text{Attention}(query = X_n, value = X_g) + X_g. \quad (6)$$

4.3 Prediction & Training

By inputting the final grid representations X_g^f into feed-forward layers, we get the predicted grid carbon emission $\hat{Y}$. During training, we employ the well-acknowledged mean absolute error loss as the prediction loss function, which can be formulated as

$$\ell_{p,i} = ||\hat{y}_i - y_i||. \quad (7)$$

The overall loss function is a weighted addition of the contrastive learning loss and the prediction loss:

$$\mathcal{L} = \sum_{i \in \mathcal{A}} \ell_{p,i} + \alpha \cdot \ell_{c,i}. \quad (8)$$

where $\mathcal{A}$ is the grid set of our target region and α is a tunable coefficient. Notably, in implementation, we add the contrastive learning loss during training after 100 epochs to prevent information collapse between the two modalities in the early stages of training.

5 Experimental Results

5.1 Experimental Setups

Datasets

We select three representative regions with varying levels of development for a comprehensive evaluation: Greater London (UK), Beijing (China), and Yinchuan (China). The first two regions are relatively well-developed, yet they exhibit distinct industrial structures and emission levels. In contrast, Yinchuan is a relatively underdeveloped region. The sources and basic statistics of the three datasets are provided in Table 1. The grid-level carbon emission statistics are all collected from ODIAC [Oda *et al.*, 2018], a well-acknowledged carbon emission inventory constructed using the hybrid means of point emission calculations and top-down allocations. Consistent with ODIAC, we set our prediction resolution as $1km \times 1km$.

Region	Great London	Beijing	Yinchuan
Area	$778\,km^2$	$1381\,km^2$	$475\,km^2$
GDP pc ($)	$71k$	$27k$	$12k$
POI Source	SafeGraph	Map Service	Map Service
Target Year	2018	2018	2019

Table 1: Summary statistics of our three main datasets.

Baselines

We compare with both carbon emission prediction methods and satellite image-based socioeconomic indicator prediction methods. The carbon emission prediction methods typically incorporate related statistics to predict regional carbon emissions and we adapt them to the high-resolution grid level:

- **SVM** [Mladenović *et al.*, 2016]. Support vector machine method for emission prediction.
- **Stacked-RFR** [Zhang *et al.*, 2022]. An ensemble two-layer stacked random forest regression model.
- **BPNN** [Zhang *et al.*, 2021]. A classical back propagation neural network.
- **CarbonGCN** [Chen *et al.*, 2024]. A graph neural network for neighborhood interaction modeling.

Since the inputs of these methods are vector-based, we transform satellite images into vectors by an acknowledged pretrained-encoder [Han *et al.*, 2020] and concatenate them with the POI count vector as input. Satellite image-based socioeconomic indicator prediction methods include:

- **READ**[Han *et al.*, 2020]: A pretrained satellite image model using transfer learning on a large-scale partially-labeled dataset to learn robust and lightweight representations.
- **Tile2Vec**[Jean *et al.*, 2019]: Models the first law of geography by using triplet loss to maximize the similarity of geo-adjacent satellite image representations.
- **PG-SimCLR** [Xi *et al.*, 2022]: Encourages geo-adjacent grids with similar facility distributions to have similar representations.

We adapt the unsupervised Tile2Vec and self-supervised PG-SimCLR methods into a supervised approach by adding a carbon emission prediction loss term and balancing the multiple losses. We also incorporate POI count vectors at the prediction stage for fair comparison.

Metrics and Implementation

To measure the prediction performance, we adopt three commonly used evaluation metrics: mean absolute error (MAE), rooted mean squared error (RMSE), and coefficient of determination (R^2) [Xi *et al.*, 2022]. We performed a grid search on our hyperparameters, including learning rate, batch size, and neighborhood scale, and balancing coefficient. The search range for the neighborhood scale is {3, 5, 7}, and that for the balancing coefficient is {1e-1, 1e-2, 1e-3}. The hyperparameters of the baselines are all carefully grid searched. During training, we partition the dataset by **regional divisions**, selecting certain sub-regions for training and validation while reserving the remaining sub-regions for testing. As an example, for the Beijing dataset, we perform training and validation

Groups	Models	London			Beijing			Yinchuan		
		R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE
Carbon Prediction	SVR	0.395	0.312	0.389	0.460	0.396	0.513	-0.167	0.664	0.803
	Stacked RFR	0.310	0.328	0.415	0.476	0.399	0.505	0.014	0.611	0.738
	BPNN	-13.315	0.456	1.890	0.080	0.466	0.669	-0.668	0.821	0.960
	CarbonGCN	-0.152	0.448	0.536	-0.082	0.579	0.726	-1.632	1.070	1.206
Satellite Image Representation	READ	0.338	0.328	0.406	0.558	0.364	0.464	-0.095	0.633	0.778
	Tile2Vec	0.527	0.269	0.343	0.567	0.352	0.459	0.070	0.598	0.717
	PG-SimCLR	0.612	0.237	0.311	0.621	0.334	0.429	0.444	0.440	0.554
	OpenCarbon	**0.786**	**0.168**	**0.231**	**0.691**	**0.302**	**0.388**	**0.622**	**0.357**	**0.457**
	Improv.	**28.4%**	**29.1%**	**25.8%**	**11.3%**	**9.6%**	**9.7%**	**40.1%**	**18.6%**	**17.4%**

Table 2: Performance comparisons on three main datasets. Training and testing datasets are split by regional divisions. The best results are in bold and the second-best results are underlined.

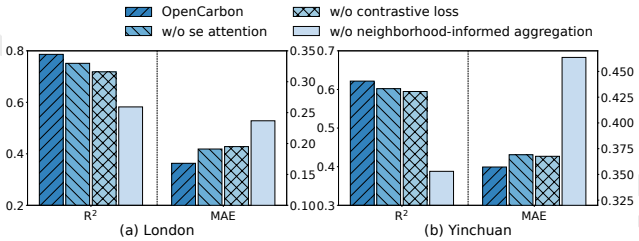

Figure 3: Ablation results on (a) London and (b) Yinchuan dataset.

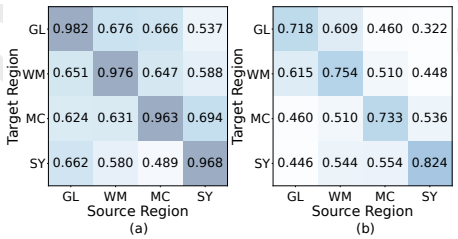

Figure 4: The Spearman's rank correlation coefficients for generalizability test on (a) OpenCarbon and (b) PG-SimCLR on Great London (GL), West Midlands (WM), Manchester Cities (MC), and South Yorkshire (SY).

on five districts (including Xicheng and Chaoyang) and conduct testing on the Fengtai district. We perform experiment on an NVIDIA RTX 4090, with 12-20 GB GPU memory usage and < 4 hour training time.

5.2 Overall Performance

We conduct extensive experiments on three distinct datasets to comprehensively evaluate the performance of OpenCarbon. From Table 2, we can draw several insights:

- **Our model's superior performance across cities with varying levels of development.** As shown in the table, our model attains the highest performance across all three datasets, surpassing the best baseline by 28.4%, 11.3%, and 40.1% in R^2 for Great London, Beijing, and Yinchuan, respectively. Across the three datasets, OpenCarbon attains an average R^2 of 0.6997—a level that is highly satisfactory for practical applications. These improvements demonstrate the consistent superiority of our model, regardless of the region or the data source.
- **Insufficient ability of existing methods in utilizing multi-modality open data to capture carbon emission's unique characteristics.** Existing carbon emission prediction methods struggle to utilize satellite images, missing their rich information. Meanwhile, current image-based approaches fail to effectively combine POI data, overlooking the complementary value of both modalities in modeling urban functionality. As a result, they cannot fully capture the functional impact on carbon emissions. Moreover, CarbonGCN's poor performance highlights the limitations of graph structures in representing the continuous layout of urban functions—a challenge our neighborhood-informed agglomeration module addresses effectively.

5.3 Ablation Study

To precisely the contribution of each design in OpenCarbon, we conduct an ablation study on one developed region, London, and one developing region, Yinchuan. Specifically, we remove the see attention and contrastive loss design in the cross-modality information extraction module, together with the neighborhood-informed aggregation module respectively. As shown in Fig 3, removing the SE attention module brings about a performance decrease of 3.8% on R^2, indicating the necessity to consider the distinct importance of different categories of facilities. Also, removing the contrastive loss design decreases R^2 by 6.49% and increases MAE by 9.49% on average. Such a performance decrease proves the essential contribution of the contrastive loss mechanism in extracting the complementary functional semantics from both the satellite image view and the facility distribution view. Further ablation studies on the neighborhood-informed aggregation module, which brings a 31.8% decrease of R^2, reveal the significant contribution of the module to characterize the agglomeration effect of emission distributions.

5.4 Generalizability Study

Unlike common socioeconomic indicators, emission factors vary across regions due to differences in energy structure and accessibility, making the same activity in different areas cause

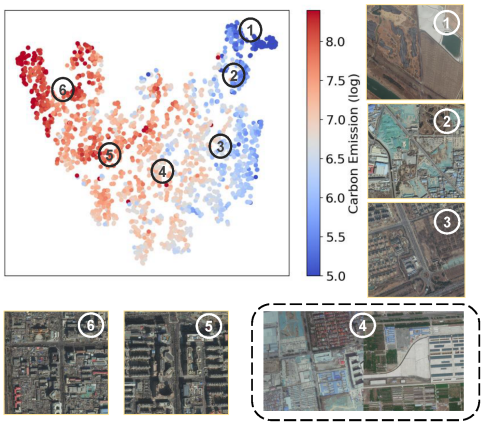

Figure 5: Visualization of the representation space. The color corresponds to the (logarithmic) grid carbon emission.

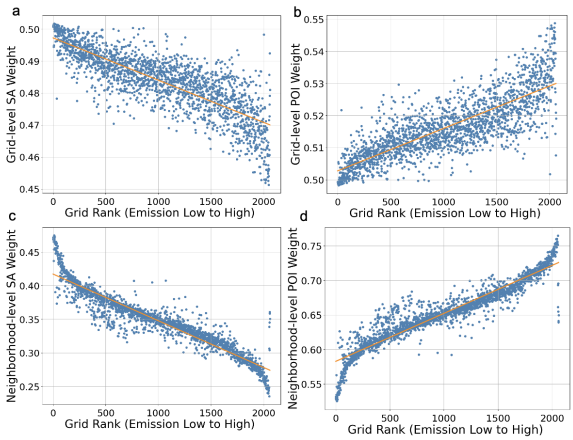

Figure 6: Visualizations of the aggregate-attention weights of the two modalities at the grid and neighborhood level. 'SA' is short for satellite images.

different carbon emissions. Therefore, there is an inherent distribution shift in carbon emissions that hinders the generalizability of prediction techniques. However, our experimental analysis shows that despite this distributional deviation, OpenCarbon's prediction results maintain their internal rank order, allowing governors to identify high-emission areas with significant mitigation potential. To evaluate our model's ranking ability, we perform direct transfer experiments: transfer the trained model to the new region directly without further fine-tuning, and use Spearman's Rank Correlation Coefficient to assess the consistency of emission magnitudes across grids. Since POI data from our three main datasets are not category-aligned due to differences in their collection sources, we did not perform generalizability across our main datasets but instead collected data on four ceremonial counties in England for testing. As shown in Figure 3, OpenCarbon performed well across all twelve source-target pairs, achieving an average Spearman's rank correlation coefficient of 0.6204, outperforming PG-SimCLR by 26.56%. This demonstrates OpenCarbon's strong generalizability in ranking emissions, highlighting its effectiveness in modeling the relationship between land use, activities, and carbon emissions.

5.5 Case Study

Grid Representation Visualization v.s. Carbon Emission

This section examines the distribution of grids with varying carbon emissions in the representation space. We use t-SNE [Van der Maaten and Hinton, 2008] to project the representations from Equation 5 into 2D, with dot colors indicating carbon emission levels. The direction in the representation space corresponds to increasing emissions. We select six anchor points along this direction and show a representative satellite image for each in Figure 5. Carbon emission levels generally correlate with land use type and density. Grids near anchor point 1 represent underdeveloped farmland and forests, while those near anchor point 2 correspond to rural areas with basic infrastructure. Anchor point 3 shows denser structures and roads, anchor point 4 includes residential and industrial zones, anchor point 5 covers high-density urban areas, and anchor point 6 highlights highly urbanized, non-residential areas. These observations highlight the link between carbon emissions and grid functionality, emphasizing the importance of incorporating functional effects into our design.

Explanatory Insights

In OpenCarbon, we incorporate aggregate-attention to aggregate the representations of the satellite image and POI modalities. Since the aggregate-attention mechanism assigns softmaxed weights to these two representations and performs a weighted sum, the change in weight during emission rise reflects the relative importance of each modality. As shown in Figure 6, for grids with higher emissions, the weights for POI modalities consistently increase at both the grid level and the neighborhood level. This interesting phenomenon suggests that fine-grained functional information becomes more important for emission prediction modeling of high-emission areas, which tend to have a higher concentration of functional activities. Furthermore, one insight from this is that, while the overall layout of the city remains largely static and difficult to change, carbon emission mitigation efforts could focus on planning specific functional activities within these layouts, steering them toward more carbon-friendly options.

6 Conclusion and Future Work

In this paper, we introduce OpenCarbon, a neighborhood-informed attentive neural network with contrastive learning for cross-modality fusion, designed to predict high-resolution carbon emissions using open data from satellite images and POI distribution. By capturing the unique functional and spatial agglomeration effects of high-resolution carbon emissions, our model significantly outperforms existing methods for both carbon emission prediction and open data-based socioeconomic forecasting. Once trained, OpenCarbon can efficiently predict carbon emissions for new locations based on the corresponding open data inputs, greatly reducing the data collection burden of traditional carbon accounting, thereby enabling targeted carbon governance and mitigation planning, and driving progress toward urban sustainability.

Acknowledgements

This work was supported by the National Key Research and Development Program of China under 2024YFC3307603. This work is also supported in part by Tsinghua University-Toyota Research Center. We also want to express gratitude to J. Zhao and S. Li for providing some processed data.

References

[Agreement, 2015] Paris Agreement. Paris agreement. In *Report of the Conference of the Parties to the United Nations Framework Convention on Climate Change (21st Session, 2015: Paris). Retrived December*, volume 4, page 2017. HeinOnline, 2015.

[Aiken *et al.*, 2022] Emily Aiken, Suzanne Bellue, Dean Karlan, Chris Udry, and Joshua E Blumenstock. Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603(7903):864–870, 2022.

[Bank, 2021] The World Bank. Advancing climate action and resilience through an urban lens, 2021.

[Böhringer and Rutherford, 2008] Christoph Böhringer and Thomas F Rutherford. Combining bottom-up and top-down. *Energy Economics*, 30(2):574–596, 2008.

[Böhringer, 1998] Christoph Böhringer. The synthesis of bottom-up and top-down in energy policy modeling. *Energy economics*, 20(3):233–248, 1998.

[Cai *et al.*, 2018] Bofeng Cai, Sai Liang, Jiong Zhou, Jinnan Wang, Libin Cao, Shen Qu, Ming Xu, and Zhifeng Yang. China high resolution emission database (chred) with point emission sources, gridded emission data, and supplementary socioeconomic data. *Resources, Conservation and Recycling*, 129:232–239, 2018.

[Campbell and Wynne, 2011] James B Campbell and Randolph H Wynne. *Introduction to remote sensing*. Guilford Press, 2011.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, 2020.

[Chen *et al.*, 2024] Yixiang Chen, Yuxin Xie, Xu Dang, Bo Huang, Chao Wu, and Donglai Jiao. Spatiotemporal prediction of carbon emissions using a hybrid deep learning model considering temporal and spatial correlations. *Environmental Modelling & Software*, 172:105937, 2024.

[Dai *et al.*, 2016] Hancheng Dai, Peggy Mischke, Xuxuan Xie, Yang Xie, and Toshihiko Masui. Closing the gap? top-down versus bottom-up projections of china's regional energy use and co2 emissions. *Applied Energy*, 162:1355–1373, 2016.

[Dhakal, 2009] Shobhakar Dhakal. Urban energy use and carbon emissions from cities in china and policy implications. *Energy policy*, 37(11):4208–4219, 2009.

[Fan *et al.*, 2018] Dongwan Fan, Kun Qin, and Chaogui Kang. Understanding urban functionality from poi space. In *2018 26th International Conference on Geoinformatics*, pages 1–6. IEEE, 2018.

[Gurney *et al.*, 2009] Kevin R Gurney, Daniel L Mendoza, Yuyu Zhou, Marc L Fischer, Chris C Miller, Sarath Geethakumar, and Stephane de la Rue du Can. High resolution fossil fuel combustion co2 emission fluxes for the united states. *Environmental science & technology*, 43(14):5535–5541, 2009.

[Gurney *et al.*, 2020] Kevin R Gurney, Jianming Liang, Risa Patarasuk, Yang Song, Jianhua Huang, and Geoffrey Roest. The vulcan version 3.0 high-resolution fossil fuel co2 emissions for the united states. *Journal of Geophysical Research: Atmospheres*, 125(19):e2020JD032974, 2020.

[Han *et al.*, 2018] Feng Han, Rui Xie, Jiayu Fang, Yu Liu, et al. The effects of urban agglomeration economies on carbon emissions: Evidence from chinese cities. *Journal of Cleaner Production*, 172:1096–1110, 2018.

[Han *et al.*, 2020] Sungwon Han, Donghyun Ahn, Hyunji Cha, Jeasurk Yang, Sungwon Park, and Meeyoung Cha. Lightweight and robust representation of economic scales from satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 428–436, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[Hutchins *et al.*, 2017] Maya G Hutchins, Jeffrey D Colby, Gregg Marland, and Eric Marland. A comparison of five high-resolution spatially-explicit, fossil-fuel, carbon dioxide emission inventories for the united states. *Mitigation and Adaptation Strategies for Global Change*, 22(6):947–972, 2017.

[Jean *et al.*, 2019] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3967–3974, 2019.

[Liu *et al.*, 2020] Zhu Liu, Philippe Ciais, Zhu Deng, Steven J Davis, Bo Zheng, Yilong Wang, Duo Cui, Biqing Zhu, Xinyu Dou, Piyu Ke, et al. Carbon monitor, a near-real-time daily dataset of global co2 emission from fossil fuel and cement production. *Scientific data*, 7(1):392, 2020.

[Lu *et al.*, 2017] Xiangyong Lu, Kaoru Ota, Mianxiong Dong, Chen Yu, and Hai Jin. Predicting transportation carbon emission with urban big data. *IEEE Transactions on Sustainable Computing*, 2(4):333–344, 2017.

[Mladenović *et al.*, 2016] Igor Mladenović, Svetlana Sokolov-Mladenović, Milos Milovančević, Dušan Marković, and Nenad Simeunović. Management and estimation of thermal comfort, carbon dioxide emission and

economic growth by support vector machine. *Renewable and sustainable energy reviews*, 64:466–476, 2016.

[Nejat *et al.*, 2015] Payam Nejat, Fatemeh Jomehzadeh, Mohammad Mahdi Taheri, Mohammad Gohari, and Muhd Zaimi Abd Majid. A global review of energy consumption, co2 emissions and policy in the residential sector (with an overview of the top ten co2 emitting countries). *Renewable and sustainable energy reviews*, 43:843–862, 2015.

[Oda *et al.*, 2018] Tomohiro Oda, Shamil Maksyutov, and Robert J Andres. The open-source data inventory for anthropogenic co2, version 2016 (odiac2016): a global monthly fossil fuel co2 gridded emissions data product for tracer transport simulations and surface flux inversions. *Earth System Science Data*, 10(1):87–107, 2018.

[Olivier *et al.*, 1994] JGJ Olivier, AF Bouwman, CWM Van der Maas, and JJM Berdowski. Emission database for global atmospheric research (edgar). *Environmental Monitoring and Assessment*, 31:93–106, 1994.

[Olivier *et al.*, 1999] Jos GJ Olivier, Jan Pieter J Bloos, Jan JM Berdowski, Antoon JH Visschedijk, and Alex F Bouwman. A 1990 global emission inventory of anthropogenic sources of carbon monoxide on 1×1 developed in the framework of edgar/geia. *Chemosphere-Global Change Science*, 1(1-3):1–17, 1999.

[Perez *et al.*, 2017] Anthony Perez, Christopher Yeh, George Azzari, Marshall Burke, David Lobell, and Stefano Ermon. Poverty prediction with public landsat 7 satellite imagery and machine learning. *arXiv preprint arXiv:1711.03654*, 2017.

[Stechemesser and Guenther, 2012] Kristin Stechemesser and Edeltraud Guenther. Carbon accounting: a systematic literature review. *Journal of Cleaner Production*, 36:17–38, 2012.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Wang *et al.*, 2015] Zijia Wang, Feng Chen, and Taku Fujiyama. Carbon emission from urban passenger transportation in beijing. *Transportation Research Part D: Transport and Environment*, 41:217–227, 2015.

[Wang *et al.*, 2016] Hongjian Wang, Daniel Kifer, Corina Graif, and Zhenhui Li. Crime rate inference with big data. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 635–644, 2016.

[Wang *et al.*, 2018] Bo Wang, Yefei Sun, and Zhaohua Wang. Agglomeration effect of co2 emissions and emissions reduction effect of technology: A spatial econometric perspective based on china's province-level data. *Journal of cleaner production*, 204:96–106, 2018.

[Wang *et al.*, 2020] Feng Wang, Wenna Fan, Juan Liu, Ge Wang, and Wei Chai. The effect of urbanization and spatial agglomeration on carbon emissions in urban agglomeration. *Environmental science and pollution research*, 27:24329–24341, 2020.

[Wu *et al.*, 2022] Xiaomeng Wu, Daoyuan Yang, Ruoxi Wu, Jiajun Gu, Yifan Wen, Shaojun Zhang, Rui Wu, Renjie Wang, Honglei Xu, K Max Zhang, et al. High-resolution mapping of regional traffic emissions using land-use machine learning models. *Atmospheric Chemistry and Physics*, 22(3):1939–1950, 2022.

[Xi *et al.*, 2022] Yanxin Xi, Tong Li, Huandong Wang, Yong Li, Sasu Tarkoma, and Pan Hui. Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests. In *Proceedings of the ACM Web Conference 2022*, pages 3308–3316, 2022.

[Xu *et al.*, 2020] Yanan Xu, Yanyan Shen, Yanmin Zhu, and Jiadi Yu. Ar2net: An attentive neural approach for business location selection with satellite data and urban data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(2):1–28, 2020.

[Yang *et al.*, 2020] Di Yang, Weixin Luan, Lu Qiao, and Mahardhika Pratama. Modeling and spatio-temporal analysis of city-level carbon emissions based on nighttime light satellite imagery. *Applied Energy*, 268:114696, 2020.

[Yeh *et al.*, 2020] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):1–11, 2020.

[Yu *et al.*, 2018] Donghan Yu, Yong Li, Fengli Xu, Pengyu Zhang, and Vassilis Kostakos. Smartphone app usage prediction using points of interest. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–21, 2018.

[Zhang *et al.*, 2015] Xiaoling Zhang, Lizi Luo, and Martin Skitmore. Household carbon emission research: an analytical review of measurement, influencing factors and mitigation prospects. *Journal of Cleaner Production*, 103:873–883, 2015.

[Zhang *et al.*, 2021] Xiaoping Zhang, Fengying Yan, Hongjiang Liu, and Zhi Qiao. Towards low carbon cities: A machine learning method for predicting urban blocks carbon emissions (ubce) based on built environment factors (bef) in changxing city, china. *Sustainable Cities and Society*, 69:102875, 2021.

[Zhang *et al.*, 2022] Yucong Zhang, Xinjie Liu, Liping Lei, and Liangyun Liu. Estimating global anthropogenic co2 gridded emissions using a data-driven stacked random forest regression model. *Remote Sensing*, 14(16):3899, 2022.

[Zhao *et al.*, 2022] Wufan Zhao, Mengmeng Li, Cai Wu, Wen Zhou, and Guozhong Chu. Identifying urban functional regions from high-resolution satellite images using a context-aware segmentation network. *Remote Sensing*, 14(16):3996, 2022.

[Zhou *et al.*, 2022] Qi Zhou, Yuheng Zhang, Ke Chang, and Maria Antonia Brovelli. Assessing osm building completeness for almost 13,000 cities globally. *International Journal of Digital Earth*, 15(1):2400–2421, 2022.