

AirRadar: Inferring Nationwide Air Quality in China with Deep Neural Networks

Qiongyan Wang¹, Yutong Xia², Siru Zhong¹, Weichuang Li¹, Yuankai Wu³,
Shifen Cheng⁵, Junbo Zhang⁴, Yu Zheng⁴, Yuxuan Liang^{1,5*}

¹The Hong Kong University of Science and Technology (Guangzhou)

²National University of Singapore

³Sichuan University

⁴JD Intelligent Cities Research

⁵State Key Lab of Resources and Environmental Information System, Chinese Academy of Sciences

{qiongyanwang, yutong.x, siruzhong, msjunbozhang, msyuzheng, yuxliang}@outlook.com;

wli043@connect.hkust-gz.edu.cn; wuyk0@scu.edu.cn; chengsf@reis.ac.cn

Abstract

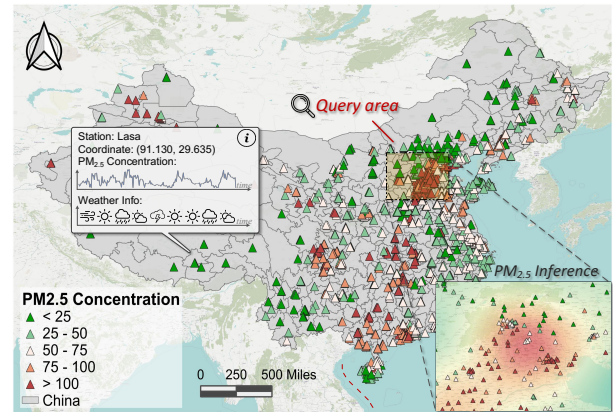
Monitoring real-time air quality is essential for safeguarding public health and fostering social progress. However, the widespread deployment of air quality monitoring stations is constrained by their significant costs. To address this limitation, we introduce *AirRadar*, a deep neural network designed to accurately infer real-time air quality in locations lacking monitoring stations by utilizing data from existing ones. By leveraging learnable mask tokens, *AirRadar* reconstructs air quality features in unmonitored regions. Specifically, it operates in two stages: first capturing spatial correlations and then adjusting for distribution shifts. We validate *AirRadar*'s efficacy using a year-long dataset from 1,085 monitoring stations across China, demonstrating its superiority over multiple baselines, even with varying degrees of unobserved data.

Introduction

Air pollution poses a significant global health threat, causing numerous premature deaths annually. Pollutants such as CO, SO₂, and PM_{2.5} continuously contaminate the air we breathe, exceeding the World Health Organization (WHO) standards for over 90% of the global population (Vallero 2014). This widespread problem necessitates urgent and comprehensive solutions for social welfare, aligning with the Sustainable Development Goals (Rafaj et al. 2018).

Given this context, monitoring real-time air quality is crucial for informing effective interventions and policies aimed at safeguarding both public health and the environment. However, the substantial costs associated with establishing monitoring stations, often exceeding 0.2 million USD per station (Zheng, Liu, and Hsieh 2013), along with their limited quantity and distribution, present a significant challenge to achieving comprehensive data coverage.

Over the past decade, *air quality inference* (Zheng, Liu, and Hsieh 2013) has focused on estimating real-time air quality in areas lacking sensors using data from existing ones. Early studies utilized non-parametric methods like k-Nearest Neighbors (Cover 1968) and Kriging (Oliver and



(a) Spatial distribution of 1,085 air quality monitoring stations throughout the Chinese Mainland. Query area: The unknown area we need to infer the air quality

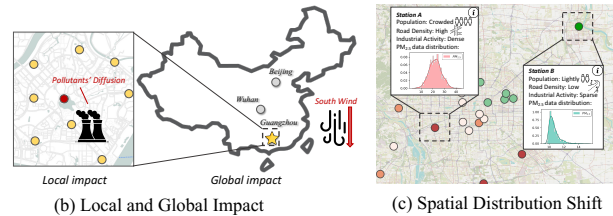


Figure 1: Illustration of air quality inference.

Webster 1990). With deep learning advances, DL-based model have become prominent for capturing complex spatial correlations among monitoring stations (Cheng et al. 2018; Han, Lu, and Chen 2021; Hu et al. 2023; Cheng et al. 2024). However, these models are mostly limited to small-scale tasks at the city level.

This work focuses on inferring air quality across the *Chinese Mainland* using thousands of monitoring stations, covering both densely monitored and sparsely monitored areas (Figure 1). This broad coverage enhances public value, has high social impact, and provides richer data for model training (Zhao et al. 2015).

Nationwide air quality inference faces challenges in capturing *complex spatial correlations*, as per the first law of ge-

*Corresponding author. Email: yuxliang@outlook.com
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

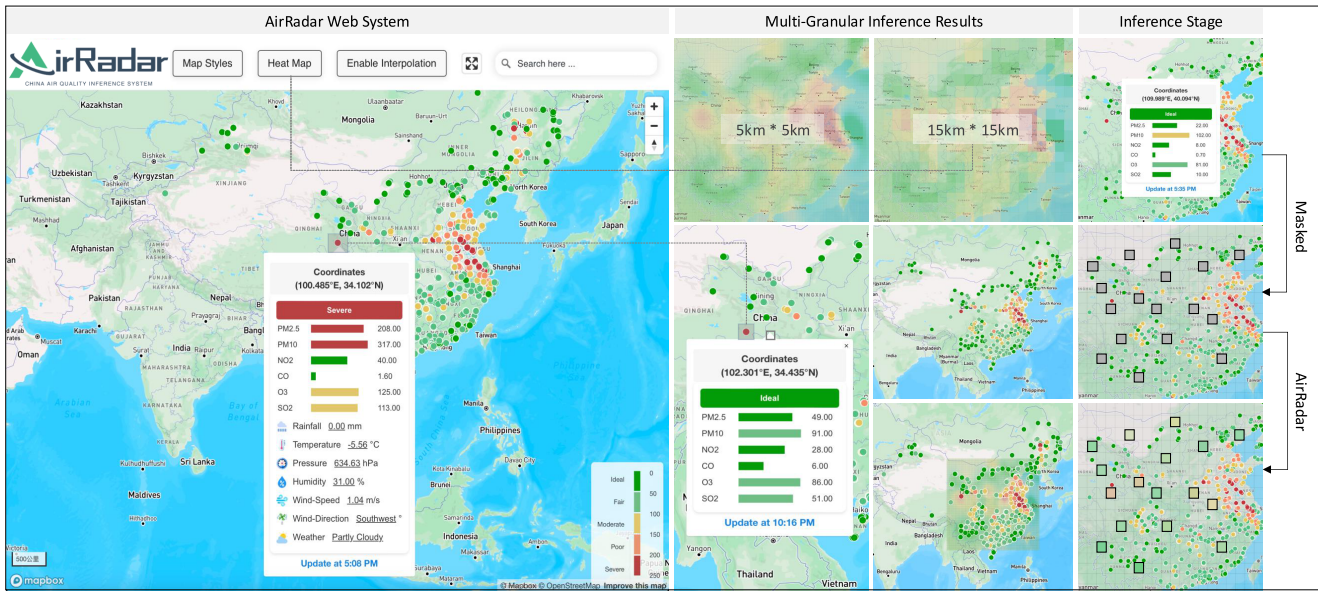


Figure 2: The user interface of our system. It can produce multi-granular inference results.

ography (Waters 2017). Nearby stations influence each other more, while distant sources also impact stations due to pollutant diffusion (Figure 1(b)). Modeling such global spatial relationships using deep GNN layers (Kipf and Welling 2016; Zhang et al. 2024) or self-attention mechanisms (Liang et al. 2023) becomes computationally prohibitive with datasets covering thousands of stations.

The second challenge arises from *spatial heterogeneity*, aligned with the second law of geography (Tobler 1970). As shown in Figure 1(c), the concentration of air pollutants is impacted by various region-specific factors, also referred to as context. Unlike city-level inference, nationwide inference must account for diverse regional contexts, making it essential to address spatial heterogeneity for accurate predictions.

To address these issues, we present a new DL-based approach called **AirRadar**. In unobserved regions – akin to areas covered in fog – there exists a critical need for a model that can function like a radar, detecting air quality where direct measurements are unavailable. AirRadar serves this purpose by providing essential air quality data in places lacking direct observation, effectively cutting through the “fog” of uncertainty. To enhance this capability, we employ mask tokens to indicate unobserved data, allowing the model to identify and address gaps in information more accurately.

Targeting the first challenge, we propose a *Spatial Learning Module* that captures spatial correlations locally and globally. The local learner utilizes dartboard projection (Liang et al. 2023) to effectively process the local information. The global learner uses the Fast Fourier Transform (FFT) to transform the feature into the frequency domain and capture global relationships with less computational complexity. For the second challenge, our *Causal Learning Module* applies causal theory to interpret causal-effect relationships in real-world contexts, dynamically weighting each context based on its contribution.

- We introduce a novel neural network named AirRadar, leveraging both local and global spatial learners to effectively capture complex spatial relationships. AirRadar facilitates air quality inference through masked feature reconstruction, allowing us to utilize a limited dataset to infer air quality in the majority of regions.
- We take a causal look and employ backdoor adjustment to address the spatial heterogeneity. Our adaptive weight structure allocates varying weights to the spatial context representations, enabling our model’s ability to effectively generalize on data from diverse, unseen spatial contexts.
- We evaluate air quality data from thousands of locations under different masking ratios, demonstrating a substantial error reduction ranging from 28.0%~44.8% compared to the state-of-the-art method STFNN. A web-based platform is deployed to verify its practicality (see Figure 2).

Preliminary

Problem Formulation

We represent current air quality readings as $\mathbf{X} \in \mathbb{R}^{N \times D}$, where N is the number of monitoring stations and D the number of measurements, including air quality and meteorological data. Historical data is denoted as $\mathbf{P} \in \mathbb{R}^{N \times T \times D}$, with T representing past time points. A set of nodes V includes observed and unobserved stations, where unobserved nodes lack readings. Given current and historical readings of observed nodes, the goal is to learn a function f to infer air quality at unobserved nodes:

$$\mathbf{X}' = f(\tilde{\mathbf{X}}, \tilde{\mathbf{P}}) \quad (1)$$

where $\tilde{\cdot}$ represents masking. \mathbf{X}' denotes the inferred results for all nodes, with $\text{PM}_{2.5}$ as the primary pollutant.

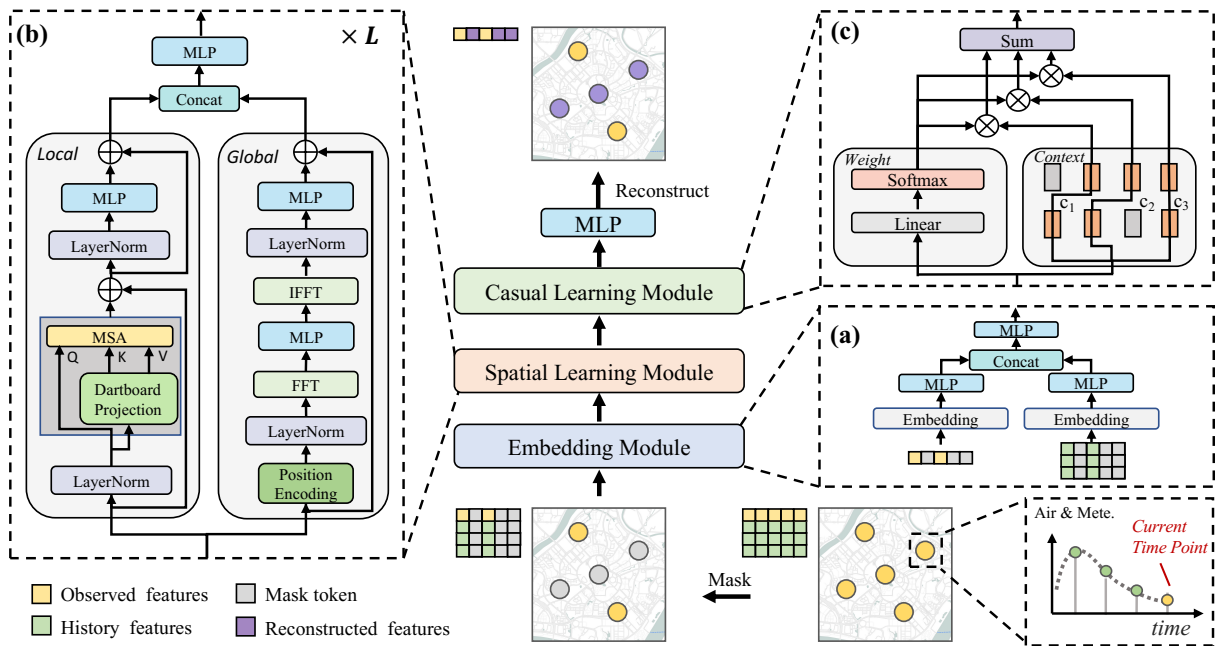


Figure 3: The framework of AirRadar. MLP: multi-layer perception. MSA: multi-head self-attention.

Related Works

Air Quality Inference. Accurately inferring air quality is paramount for societal well-being. As air quality research progresses, models integrate multi-source data to make detailed inferences. Traditional methods (Fawagreh, Gaber, and Elyan 2014; Cheng, Peng, and Lu 2020; Cheng et al. 2024) often rely on distance-based features, limiting their ability to capture site significance. Different from these models, certain studies (Cheng et al. 2018; Han, Lu, and Chen 2021) incorporate the attention mechanism to assign sensor weights automatically but face challenges with sparse data and providing detailed insights. Recently, STGNP (Hu et al. 2023) introduced the Graph Neural Process for spatial extrapolation tasks. STFNN (Feng et al. 2024) integrates both spatial and temporal viewpoints, utilizing fields and graphs to enhance the accuracy of air quality inference. While these models have made strides in air quality inference, they fall short in addressing two critical challenges: the *complex spatial correlations* and *spatial heterogeneity*.

Neural Networks for Spatial Learning. Neural networks have been widely applied in spatial learning, leveraging various architectures to capture spatial dependencies. GNNs (Wu et al. 2022; Li et al. 2022) are commonly used to model spatial relationships in non-Euclidean space. CNNs (Guo et al. 2019; Khosravi et al. 2020) are often employed to capture local spatial features by processing data with grid-like topologies, such as images or regular grids. Recently, Transformers (Wu et al. 2020; Hao et al. 2024) have also been adapted for spatial learning, offering an attention mechanism to capture long-range dependencies. Each of these approaches contributes uniquely to enhancing spatial learning capabilities in different tasks.

Causal Inference. Causal inference (Pearl, Glymour, and

Jewell 2016) aims to explore causal relationships between variables, promoting stable and robust learning and inference processes. The integration of deep learning techniques with causal inference has demonstrated significant potential in spatial-temporal mining tasks (Xia et al. 2024; Luo et al. 2024). In this study, we apply causal methods to address spatial heterogeneity in the air quality inference task.

Methodology

Figure 3 depicts the framework of AirRadar, which consists of input masking and three learning modules. To address the issue of the target node lacking prior information, a shared learnable token is employed to mask the features of target nodes. During training, stations are masked as target nodes, while areas lacking stations serve as target nodes during inference (See Figure 2 Right). We subsequently process the masked inputs with the following modules:

- *Embedding module:* We utilize a fully-connected layer to transform the masked air quality data into feature space and integrate observed features with historical features to obtain an initial representation for each node.
- *Spatial Learning Module:* This module integrates spatial correlation from both local and global perspectives. The local part uses dartboard projection to efficiently incorporate local spatial information, while the global part transforms the feature from the spatial domain to the frequency domain to effectively capture global relationships.
- *Causal Learning Module:* We begin by constructing a structural causal model to illustrate how various contexts influence air quality. We then implement backdoor adjustment by dynamically allocating weights to each context, effectively mitigating confounding effects.

Embedding Module

Considering that observed data and historical data can provide essential information for inferring the air quality of unobserved nodes, we have taken these data into our input. However, the target nodes are without any prior information. Inspired by MAE (He et al. 2022), we employ a shared learnable token \mathbf{o} to mask the data of target nodes. Then we transform the masked data into feature space, resulting in $\hat{\mathbf{X}} \in \mathbb{R}^{N \times E}$ and $\hat{\mathbf{P}} \in \mathbb{R}^{N \times E}$. Finally, we combine the observation and historical data to create an initial spatial representation for each node, denoted as $\mathbf{H} \in \mathbb{R}^{N \times 2E}$.

Spatial Learning Module

We introduce the Spatial Learning Module, a novel spatial learner designed to capture spatial correlations from local and global perspectives based on the principle of spatial auto-correlation. Our Spatial Learning Module boasts a substantial receptive field with computational complexity less than quadratic. Figure 3b outlines its pipeline. Each block (indexed by l) takes the hidden state \mathbf{Z}^{l-1} as input, with \mathbf{H} serving as the initial embedding \mathbf{Z}^0 . This input is split into two parts: the local module (left side) and the global module (right side). In the subsequent parts of this section, we will detail these two key parts respectively.

Local Spatial Correlation. Beyond local emissions, a place’s air quality is also affected by neighboring areas. Directly using standard MSA will result in a square increase in costs, thus our objective is find a new way to *efficiently* capture local spatial correlations. We drew inspiration from (Liang et al. 2023) and used a dartboard projection to achieve this. As depicted in Figure 4, node i ’s local range is defined using a dartboard, which divides its neighborhood space into several distinct regions. Nodes positioned beyond the outermost circle fall outside the local range. Different from (Liang et al. 2023), we modify the orientation of the central angle to align with the potential wind direction, thereby enhancing integration with external factors such as wind direction and the interpretability of the model.

The latent features \mathbf{Z}^{l-1} are used to generate the query vector \mathbf{Q}_h for each node. For the query nodes, the key vector and value vector are obtained via dartboard projection. The dartboard projection matrix for each node i is represented as $\mathbf{M}_i \in \mathbb{R}^{G \times N}$, indicating how nearby nodes are mapped to G regions defined by the line fragments and circles. Here, each entry $\mathbf{m}_{js} = 1$ signifies that the s -th node belongs to the j -th region. We employ the dartboard projection matrix to project the node features into local regional representations. The projected features of each region $\mathbf{R}_i \in \mathbb{R}^{G \times 2E}$ are generated by:

$$\mathbf{R}_i = \mathbf{M}_i \mathbf{F} = \sum_{\forall j} \mathbf{m}_{ij} \mathbf{f}_j, \quad (2)$$

with projection matrix \mathbf{M}_i and normalized node features $\mathbf{f}_j \in \mathbb{R}^{2E}$. The ensuing step involves concatenating these representations after dartboard projection $\mathbf{R} = [\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_N]$. Subsequently, we feed this regional representation into linear layers to generate the key vector $\mathbf{K}_h \in \mathbb{R}^{N \times G \times 2E}$ and value vector $\mathbf{V}_h \in \mathbb{R}^{N \times G \times 2E}$. The

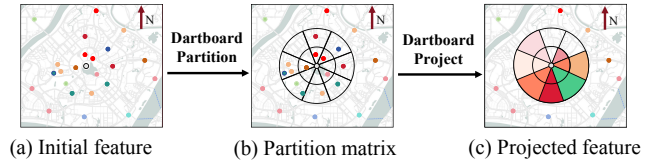


Figure 4: Dartboard projection: For a query node (gray point), its local range is partitioned into regions by circles and lines. The central angle aligns with wind directions.

implementation of multi-head self-attention efficiently captures detailed local spatial correlations between the query node and its neighboring regions, as articulated by:

$$\mathbf{X}_h = \text{Softmax}(\alpha \mathbf{Q}_h \mathbf{K}_h^T + \mathbf{B}_h) \mathbf{V}_h, \quad (3)$$

where $\mathbf{B}_h \in \mathbb{R}^{N \times G}$ is a learnable relative position encoding.

This local spatial learner integrates domain knowledge of air diffusion while addressing computational expense. With a smaller number of regions ($G \ll N$), the computational complexity, $O(NGE)$, grows linearly with the number of stations, proving more efficient than standard MSA with complexity $O(N^2E)$.

Global Spatial Correlation. In the context of air quality inference, beyond accounting for local spatial relationships, it is essential to incorporate considerations of global spatial relationships, i.e., impact from other cities.

To avoid the computational complexity growth with the quadratic cost, we consider using the Fourier Neural Operator (FNO) to reduce the computational burden, inspired by (Li et al. 2020a). We first conceptualize the process of self-attention as a continuous kernel integrals (Guibas et al. 2021; Li et al. 2020b), denoted as follows:

$$\zeta(X)(i) = \int_D \kappa(i, j) X(j) dj \quad \forall i \in \mathbb{D}, \quad (4)$$

where $\zeta(X)(i)$ is equivalent to the i -th row of the similarity matrix $Att \in \mathbb{R}^{N \times N}$.

This means that when calculating $\zeta(X)(i)$, the information on the entire domain is considered and the global information is captured. Furthermore, this operation can be efficiently implemented by FFT with computational complexity $O(NE \log N)$:

$$\zeta(X)(i) = F^{-1}(F(\kappa) \cdot F(X))(i) \quad i \in \mathbb{D}, \quad (5)$$

where \cdot denotes matrix multiplication and F, F^{-1} denote the FFT and its inverse, respectively.

We treat each node as the token in the sequence. For the input feature $\mathbf{Z}^{l-1} \in \mathbb{R}^{N \times 2E}$, we need to define the complex-valued weight tensor $\hat{\mathbf{W}}$ to parameterize the kernel. To reduce the parameter count, we impose a block diagonal structure on $\hat{\mathbf{W}}$, where it is divided into \hat{K} weight blocks of size $2E/\hat{K} \times 2E/\hat{K}$. Therefore, the global spatial mixing operations are defined as follows:

$$\begin{aligned} \mathbf{g}_i &= [\text{FFT}(\mathbf{Z}^{l-1})]_i \\ \tilde{\mathbf{g}}_i^k &= \hat{\mathbf{W}}_k \mathbf{g}_i^k + \mathbf{b}_k \quad k = 1, 2, \dots, \hat{K} \\ \hat{\mathbf{g}}_i &= [\text{IFFT}(\text{S}(\tilde{\mathbf{g}}, \lambda))]_i \end{aligned} \quad (6)$$

where $\hat{\mathbf{W}}$ are shared for all nodes, the computational complexity of node mixing is $O(NE^2/\hat{K})$. The scalar λ represents the sparsity threshold, while S denotes the soft-thresholding and shrinkage operation.

To efficiently capture global spatial relationships, we transform input features from spatial to frequency domains and employ shared weights for spatial mixing. The computational complexity of this FNO with shared weights is $O(NE^2/\hat{K} + NE \log N)$, more efficient than $O(N^2E + 3NE^2)$, especially with increasing node count. After obtaining local and global spatial representations for each node, we combined them, denoted as $\mathbf{Z}^l \in \mathbb{R}^{N \times 2E}$.

Casual Learning Module

In air quality inference, as the geographical scope widens, the challenge of spatial heterogeneity emerges. To tackle this, we leverage causality theory by establishing a Structural Causal Model (SCM) (Pearl et al. 2000). In the SCM shown in Figure 5a, C , X , and Y denote random variables for spatial context, observed data, and target area air pollutants, respectively, with arrows representing causal-effect relationships. Based on the above definitions, the causal relationships in Figure 5a can be denoted as $P(X, Y|C) = P(X|C)P(Y|X, C)$.

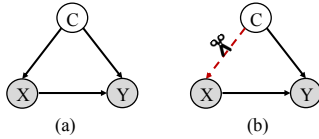


Figure 5: SCMs of (a) Data generation under Real-world Scenarios; (b) Backdoor Adjustment for C .

Confounder & Intervention. We observe that there exists a backdoor path $X \leftarrow C \rightarrow Y$ between X and Y , where context C acts as a confounding factor. However, this confounding influence of C is undesirable for accurate inference, as it introduces potential bias in the model’s inference.

The *backdoor adjustment* (Yang et al. 2022; Sui et al. 2022; Xia et al. 2024) can cut off the causal relationship from C to X to mitigate the confounding effects, as shown in the red dashed arrow in Figure 5b. Therefore, we employ a backdoor adjustment to estimate Y via $P(Y|do(X))$ by stratifying into discrete pieces $C = \{c_i\}_{i=1}^{|C|}$ as:

$$P(Y|do(X)) = \sum_{i=1}^{|C|} P(Y|X, C = c_i)P(C = c_i) \quad (7)$$

Although the backdoor path may seem effectively severed, optimizing with Eq. 7 becomes intractable due to challenges in defining c_i and the unknown prior distribution $P(C)$. To overcome this, we treat C as a latent variable obtained by a variational distribution $Q(C|X)$ (Yang et al. 2022).

Estimating this distribution is facilitated by the Casual Learning Module, illustrated in Figure 3c. Assuming that the contexts are independent can lead to overfitting issues, thus we denote a context as a 0-1 matrix, i.e.,

$$c_k = \text{stack}[c_k^1, c_k^2, \dots, c_k^U] \quad (8)$$

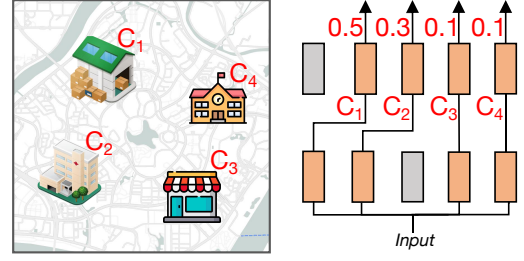


Figure 6: An explanation of how AirRadar combines the influence of each context. c_k denotes the context type.

where U denotes the number of layers and K denotes the number of contexts. Each context type corresponds to a specific combination of inference units. To comprehensively consider the influence of various contexts, we design a network that can dynamically allocate weights:

$$\mathbf{Y}^{l+1} = \sum_{k=1}^K \Psi^l(\mathbf{Y}^l)[k] \Phi_k^l(\mathbf{Y}^l), \quad (9)$$

where Ψ is the dynamic weight allocation structure and Φ is the context-specific encoder. The computation represents a realization of c drawn from the conditional distribution $Q(C|X = \mathbf{X})$. As shown in Figure 6, different contexts may be linked by sharing the same inference units at certain layers, serving as an effective inductive bias that helps the model eliminate the influence of spatial heterogeneity and generalize to unseen areas.

Inference & Optimization

Upon obtaining the final feature representation \mathbf{Y} , we employ an MLP for feature reconstruction to restore the air quality data of all nodes as:

$$\mathbf{X}' = \text{MLP}(\mathbf{Y}) \quad (10)$$

where $\mathbf{X}' \in \mathbb{R}^{N \times D}$ is the reconstrued data for all nodes. This process culminates in the completion of the target inference. For optimization, we utilize the L1 loss as the loss function \mathcal{L} , expressed as follows:

$$\mathcal{L}(\mathbf{X}, \mathbf{X}') = \frac{1}{N_u} \sum_{v_i \in \tilde{V}} |\mathbf{X}_i - \mathbf{X}'_i| \quad (11)$$

where N_u is the number of unobserved nodes. Note that we exclusively compute the loss for the target nodes.

Experiments

In this section, we conduct experiments to address the following research questions:

- **RQ1:** Can AirRadar outperform existing approaches on real-time air quality inference in the Chinese Mainland?
- **RQ2:** How does each component affect performance?
- **RQ3:** What is the effect of the major hyperparameters?
- **RQ4:** How is the practicality of AirRadar?

Model	Year	Missing Ratio = 25%				Missing Ratio = 50%				Missing Ratio = 75%			
		MAE	Δ	RMSE	MAPE	MAE	Δ	RMSE	MAPE	MAE	Δ	RMSE	MAPE
KNN	1967	30.50	+173.8%	65.40	1.36	30.25	+167.2%	72.23	0.71	34.07	+202.3%	74.55	0.64
RF	2001	29.22	+162.3%	68.95	0.76	29.71	+162.5%	71.61	0.75	29.82	+164.6%	70.99	0.74
SGNP	2019	23.60	+96.0%	37.58	0.83	24.06	+112.5%	37.08	0.93	21.68	+89.1%	33.68	0.84
STGNP	2022	23.21	+90.6%	38.13	0.62	21.95	+93.9%	37.13	0.67	19.58	+73.7%	31.95	0.69
VAE	2013	28.49	+155.7%	67.11	0.94	28.92	+155.5%	69.67	0.94	29.00	+157.3%	69.11	0.93
GAE	2016	12.63	+13.4%	23.80	0.46	12.78	+12.9%	24.11	0.46	12.57	+11.5%	23.73	0.46
GraphMAE	2022	12.40	+11.3%	23.20	0.46	12.32	+8.8%	23.11	0.46	11.59	+2.8%	21.51	0.43
MCAM	2021	23.94	+120.9%	36.25	0.95	25.01	+103.0%	37.94	0.92	25.19	+123.5%	37.82	1.04
STFNN	2024	11.14	-	23.20	0.39	11.32	-	19.91	0.41	11.27	-	19.86	0.41
AirRadar	-	6.41	-40.0%	12.60	0.24	6.79	-44.9%	12.90	0.26	8.11	-28.0%	14.84	0.29

Table 1: Model comparison on the nationwide dataset. The symbol Δ represents the reduction in MAE compared to the state-of-the-art method STFNN (Feng et al. 2024) The missing ratio represents the proportion of unobserved nodes to all nodes.

Experimental Settings

Datasets. Our system (currently anonymous) collected nationwide air quality datasets for a year (Jan. 1, 2018 to Dec. 31, 2018). Figure 1 illustrates the distribution of monitoring stations, showcasing broad spatial coverage exceeding previous datasets. Meteorological data was also incorporated to consider exogenous influences. Input features encompass continuous (e.g., pollutant concentrations, temperature) and categorical (e.g., weather conditions, wind direction) variables. Categorical features were transformed into real-valued vectors via embedding. Following precedents (Cheng et al. 2018; Hu et al. 2023), known air quality data was utilized to infer $PM_{2.5}$ values for other sites. The dataset was split into training (60%), validation (20%), and testing (20%) sets, comprising 5,256, 1,752, and 1,752 instances respectively.

Implement Details. Our model, implemented with PyTorch 2.1.0 on an NVIDIA RTX 2080Ti GPU, uses the Adam optimizer with a batch size 32, and an initial learning rate of 5×10^{-3} . The hidden size (E) is 32. We randomly mask a percentage of node features and include 2 spatial learning blocks. The Local part has 2 heads with 50km and 200km ranges. The global part comprises 2 weight blocks, and the Casual Learning Module uses 4 contexts.

Baselines & Evaluation Metrics. We compare our AirRadar with strong baselines that belong to four categories:

- **Statistical models:** **KNN** (Cover 1968) and **RF** (Fawagreh, Gaber, and Elyan 2014) are conventional baselines for spatial interpolation, e.g., air quality inference.
- **NP-based models:** We take several popular NP-based models as baselines, i.e., **SGNP** which is a modification of (Singh et al. 2019), and **STGNP** (Hu et al. 2023).
- **Autoencoder-based models:** We also take some popular Autoencoders as baselines, i.e., **VAE** (Kingma and Welling 2013), **GAE** and **GraphMAE** (Hou et al. 2022).
- **Other models:** We select a couple of other models for comparison, including **MCAM** (Han, Lu, and Chen 2021), **STFNN** (Feng et al. 2024).

To evaluate our model, we follow STFNN (Feng et al. 2024) to employ Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) as the evaluation metrics.

Model Comparison (RQ1)

To answer RQ1, we perform a model comparison in terms of MAE, RMSE, and MAPE. Table 1 presents the experimental results over the nationwide dataset. Our model significantly outperforms all competing baselines over all kinds of missing ratios. Compared to the STFNN, our approach reduces MAE by approximately 28% to 44.9% in each missing ratio. There are three primary reasons for the efficacy of our model. Firstly, our method effectively captures comprehensive spatial correlations. Secondly, AirRadar leverages backdoor adjustment to eliminate confounding effects and dynamically assigns weights to each potential context representation. Lastly, using the learnable token to represent the target node empowers robust inference even for unobserved nodes. In contrast, traditional models (KNN, RF) lack strong learning capabilities, while AE-based, NN-based, and NP-based models struggle with spatial range and diversity, limiting their adaptability for nationwide air quality inference.

Ablation Study (RQ2)

To evaluate the individual contributions of each component to our model’s overall performance and address RQ2, we conduct ablation studies. The results are presented in Figure 7 and Table 2. Each study involves modifying only the corresponding part while keeping other settings unchanged.

Effects of Historical Data. As a pivotal component of AirRadar, historical data plays a fundamental role in offering prolonged observations of the environmental background, encompassing factors such as meteorological conditions and seasonal variations. As illustrated in Figure 7, the exclusion of historical data leads to a noticeable decline in performance across all missing ratios.

Effects of Spatial Learning Module. To assess the impact of the Spatial Learning Module, we examine the following

variants for comparison: a) **w/o Local**: The local component in the Spatial Learning Module is disabled, meaning that no local spatial relationships are considered. b) **w/o Global**: The global component in the Spatial Learning Module is omitted, indicating that the global spatial correlation is not taken into account in this model. The results are shown in the Table 2. When eliminating both local and global components, we observe a significant deterioration in MAE, especially from removing the global part. This highlights the importance of considering spatial correlations from both perspectives. Second, AirRadar obtains lower inference errors overall missing ratios, while running **44%** faster than GraphMAE. This advantage indicates that our Spatial Learning Module holds significant promise as a fundamental component for capturing complex spatial dependencies.

Variant	Time/epoch	25%	50%	75%
w/o local	27	8.67	7.45	9.09
w/o global	24	8.43	9.22	11.68
GraphMAE	74	12.50	12.43	12.92
AirRadar	41	6.41	6.79	8.11

Table 2: Effects of Spatial Learning Module under different missing ratios. Time/epoch: seconds per training epoch.

Effects of Casual Learning Module. To evaluate the effectiveness of the Casual Learning Module in addressing spatial heterogeneity challenges, we compare our model with the e following variants: a) **w/o Causal**: The entire Casual Learning Module is omitted from our AirRadar. b) **w/o Adaptive**: The dynamic weight allocation mechanism is excluded from the Casual Learning Module. The outcomes are presented in Figure 7. When the causal module is omitted, the model fails to mitigate the influence of potential context and, consequently, cannot address the spatial heterogeneity issue. Absent the adaptive module, the model is unable to account for varying impacts introduced by different contexts, resulting in a performance decline.

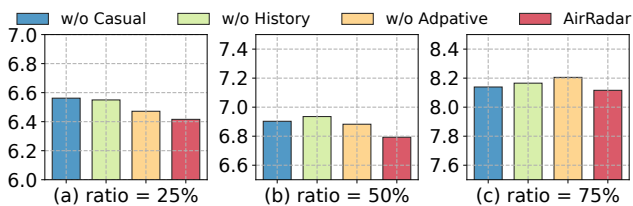


Figure 7: Effects of modules across different missing ratios.

Hyperparameter Study (RQ3)

In response to the final research question, we explore the impact of varying the hidden dimension E , the number of Spatial Learning blocks L , the number of contexts C , and the sparsity threshold λ on the MAE, with a fixed missing rate of 75%. Figure 8a presents the results of AirRadar with different values of C . Interestingly, we observe that the performance is not strictly proportional to the number of contexts, and we establish $C = 4$ as our default setting based on these

findings. We experiment with different sparsity thresholds λ ranging from 0 to 1. The MAE reaches its peak at $\lambda = 10^{-2}$, demonstrating the efficacy of the sparsity measure.

Figure 8b investigates the number of Spatial Learning blocks (L). It becomes evident that the model’s performance improves as L increases, but this improvement comes at the cost of a rapid growth in the number of parameters. Notably, when $L = 3$, we achieve a similar level of accuracy without the excessive parameter increase. Figure 8c further delves into the influence of the hidden dimension E . Here, we make two key observations: 1) A very small E (e.g., 8) results in diminished performance due to its limited capacity. 2) $E = 64$ yields similar performance, but it comes with a substantial increase in computational cost.

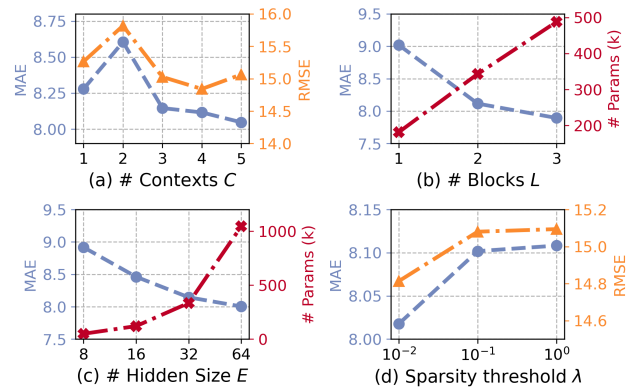


Figure 8: Hyperparameter study when missing ratio=75%.

Practicality (RQ4)

To verify the model’s practicality, we deploy a web-based system for monitoring air quality and inferring nationwide air quality in China, as shown in Figure 2. Each point on the map corresponds to an air quality monitoring station and the associated number signifies the respective air quality data at that station. The system’s practicality lies in integrating real-time data from numerous stations, providing accessible air quality information nationwide.

Conclusion & Future Work

In this paper, we introduce AirRadar for nationwide air quality inference in China. AirRadar efficiently captures intricate spatial correlations from both local and global perspectives, addressing the challenge of spatial heterogeneity through backdoor adjustment techniques with an adaptive weight mechanism. Our model can achieve state-of-the-art performance on a nationwide air quality dataset with different missing ratios. Accompanying our model, we develop a web system capable of leveraging air quality data from existing monitoring stations to infer air pollutants of regions without monitoring stations. In the future, we will consider online learning and explore the application of our model in other domains, such as traffic and climate.

Acknowledgments

This work is supported a grant from State Key Laboratory of Resources and Environmental Information System. This work is also supported by the National Natural Science Foundation of China (No. 62402414), the Guangzhou-HKUST(GZ) Joint Funding Program (No. 2024A03J0620), Guangzhou Municipal Science and Technology Project (No. 2023A03J0011), the Guangzhou Industrial Information and Intelligent Key Laboratory Project (No. 2024A03J0628), and Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007).

References

- Cheng, S.; Peng, P.; and Lu, F. 2020. A lightweight ensemble spatiotemporal interpolation model for geospatial data. *International Journal of Geographical Information Science*, 34(9): 1849–1872.
- Cheng, S.; Zhang, W.; Luo, P.; Wang, L.; and Lu, F. 2024. An explainable spatial interpolation method considering spatial stratified heterogeneity. *International Journal of Geographical Information Science*, 1–27.
- Cheng, W.; Shen, Y.; Zhu, Y.; and Huang, L. 2018. A neural attention model for urban air quality inference: Learning the weights of monitoring stations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Cover, T. 1968. Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory*, 14(1): 50–55.
- Fawagreh, K.; Gaber, M. M.; and Elyan, E. 2014. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal*, 2(1): 602–609.
- Feng, Y.; Wang, Q.; Xia, Y.; Huang, J.; Zhong, S.; and Liang, Y. 2024. Spatio-Temporal Field Neural Networks for Air Quality Inference. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 7260–7268. International Joint Conferences on Artificial Intelligence Organization. AI for Good.
- Guibas, J.; Mardani, M.; Li, Z.; Tao, A.; Anandkumar, A.; and Catanzaro, B. 2021. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*.
- Guo, S.; Lin, Y.; Li, S.; Chen, Z.; and Wan, H. 2019. Deep spatial-temporal 3D convolutional neural networks for traffic data forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 20(10): 3913–3926.
- Han, Q.; Lu, D.; and Chen, R. 2021. Fine-Grained Air Quality Inference via Multi-Channel Attention Model. In *IJCAI*, 2512–2518.
- Hao, X.; Chen, W.; Yan, Y.; Zhong, S.; Wang, K.; Wen, Q.; and Liang, Y. 2024. UrbanVLP: A Multi-Granularity Vision-Language Pre-Trained Foundation Model for Urban Indicator Prediction. *arXiv preprint arXiv:2403.16831*.
- He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.
- Hou, Z.; Liu, X.; Cen, Y.; Dong, Y.; Yang, H.; Wang, C.; and Tang, J. 2022. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 594–604.
- Hu, J.; Liang, Y.; Fan, Z.; Chen, H.; Zheng, Y.; and Zimmermann, R. 2023. Graph Neural Processes for Spatio-Temporal Extrapolation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 752–763.
- Khosravi, K.; Panahi, M.; Golkarian, A.; Keesstra, S. D.; Saco, P. M.; Bui, D. T.; and Lee, S. 2020. Convolutional neural network approach for spatial prediction of flood hazard at national scale of Iran. *Journal of Hydrology*, 591: 125552.
- Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- Li, J.; Chen, S.; Pan, X.; Yuan, Y.; and Shen, H.-B. 2022. Cell clustering for spatial transcriptomics data with graph neural networks. *Nature Computational Science*, 2(6): 399–408.
- Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; and Anandkumar, A. 2020a. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*.
- Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; and Anandkumar, A. 2020b. Neural operator: Graph kernel network for partial differential equations. *arXiv preprint arXiv:2003.03485*.
- Liang, Y.; Xia, Y.; Ke, S.; Wang, Y.; Wen, Q.; Zhang, J.; Zheng, Y.; and Zimmermann, R. 2023. Airformer: Predicting nationwide air quality in china with transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14329–14337.
- Luo, K.; Zhu, Y.; Chen, W.; Wang, K.; Zhou, Z.; Ruan, S.; and Liang, Y. 2024. Towards Robust Trajectory Representations: Isolating Environmental Confounders with Causal Learning. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 2243–2251. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Oliver, M. A.; and Webster, R. 1990. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System*, 4(3): 313–332.
- Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.
- Pearl, J.; et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2): 3.
- Rafaj, P.; Kiesewetter, G.; Gül, T.; Schöpp, W.; Cofala, J.; Klimont, Z.; Purohit, P.; Heyes, C.; Amann, M.; Borken-Kleefeld, J.; et al. 2018. Outlook for clean air in the context of sustainable development goals. *Global Environmental Change*, 53: 1–11.

- Singh, G.; Yoon, J.; Son, Y.; and Ahn, S. 2019. Sequential neural processes. *Advances in Neural Information Processing Systems*, 32.
- Sui, Y.; Wang, X.; Wu, J.; Lin, M.; He, X.; and Chua, T.-S. 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.
- Tobler, W. R. 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46(sup1): 234–240.
- Vallero, D. A. 2014. *Fundamentals of air pollution*. Academic press.
- Waters, N. 2017. Tobler’s first law of geography. *The international encyclopedia of geography*, 1–13.
- Wu, L.; Cui, P.; Pei, J.; Zhao, L.; and Guo, X. 2022. Graph neural networks: foundation, frontiers and applications. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4840–4841.
- Wu, X.; Huang, C.; Zhang, C.; and Chawla, N. V. 2020. Hierarchically structured transformer networks for fine-grained spatial event forecasting. In *Proceedings of the web conference 2020*, 2320–2330.
- Xia, Y.; Liang, Y.; Wen, H.; Liu, X.; Wang, K.; Zhou, Z.; and Zimmermann, R. 2024. Deciphering spatio-temporal graph forecasting: A causal lens and treatment. *Advances in Neural Information Processing Systems*, 36.
- Yang, C.; Wu, Q.; Wen, Q.; Zhou, Z.; Sun, L.; and Yan, J. 2022. Towards out-of-distribution sequential event prediction: A causal treatment. *Advances in neural information processing systems*, 35: 22656–22670.
- Zhang, G.; Yue, Y.; Wang, K.; Fang, J.; Sui, Y.; Wang, K.; Liang, Y.; Cheng, D.; Pan, S.; and Chen, T. 2024. Two heads are better than one: Boosting graph sparse training via semantic and topological awareness. *arXiv preprint arXiv:2402.01242*.
- Zhao, L.; Sun, Q.; Ye, J.; Chen, F.; Lu, C.-T.; and Ramakrishnan, N. 2015. Multi-task learning for spatio-temporal event forecasting. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 1503–1512.
- Zheng, Y.; Liu, F.; and Hsieh, H.-P. 2013. U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1436–1444.