atmosphere

MDPI

Article

Fine-Grained Air Pollution Inference at Large-Scale Region Level via 3D Spatiotemporal Attention Super-Resolution Model

Changqun Li [1], Shan Tang [1], Jing Liu [1], Kai Pan [2,3], Zhenyi Xu [2,3,4,5,6,*], Yunbo Zhao [3,6] and Shuchen Yang [5,*]

[1] Anhui Vocational College of Grain Engineering, Hefei 231635, China; changqun.li@foxmail.com (C.L.); 18326076518@163.com (S.T.); liujing_0621@126.com (J.L.)
[2] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, Hefei 230031, China; wa21301055@stu.ahu.edu.cn
[3] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China; ybzhao@ustc.edu.cn
[4] State Environmental Protection Key Laboratory of Vehicle Emission Control and Simulation, Chinese Research Academy of Environmental Sciences, Beijing 100012, China
[5] Jiangsu Engineering Research Center of Key Technology for Intelligent Manufacturing Equipment, Suqian University, Suqian 223800, China
[6] Institute of Advanced Technology, University of Science and Technology of China, Hefei 230088, China
* Correspondence: xuzhenyi@mail.ustc.edu.cn (Z.X.); 18146@squ.edu.cn (S.Y.)

Abstract: Air pollution presents a serious hazard to human health and the environment for the global rise in industrialization and urbanization. While fine-grained monitoring is crucial for understanding the formation and control of air pollution and their effects on human health, existing macro-regional level or ground-level methods make air pollution inference in the same spatial scale and fail to address the spatiotemporal correlations between cross-grained air pollution distribution. In this paper, we propose a 3D spatiotemporal attention super-resolution model (AirSTFM) for fine-grained air pollution inference at a large-scale region level. Firstly, we design a 3D-patch-wise self-attention convolutional module to extract the spatiotemporal features of air pollution, which aggregates both spatial and temporal information of coarse-grained air pollution and employs a sliding window to add spatial local features. Then, we propose a bidirectional optical flow feed-forward layer to extract the short-term air pollution diffusion characteristics, which can learn the temporal correlation contaminant diffusion between closeness time intervals. Finally, we construct a spatiotemporal super-resolution upsampling pretext task to model the higher-level dispersion features mapping between the coarse-grained and fined-grained air pollution distribution. The proposed method is tested on the $PM_{2.5}$ pollution datatset of the Yangtze River Delta region. Our model outperforms the second best model in RMSE, MAE, and MAPE by 2.6%, 3.05%, and 6.36% in the 100% division, and our model also outperforms the second best model in RMSE, MAE, and MAPE by 3.86%, 3.76%, and 12.18% in the 40% division, which demonstrates the applicability of our model for different data sizes. Furthermore, the comprehensive experiment results show that our proposed AirSTFM outperforms the state-of-the-art models.

Keywords: fine-grained air pollution inference; spatial–temporal features; self-attention; super-resolution

check for updates

Academic Editor: Antonio Donateo

Received: 15 December 2024
Revised: 19 January 2025
Accepted: 23 January 2025
Published: 31 January 2025

Citation: Li, C.; Tang, S.; Liu, J.; Pan, K.; Xu, Z.; Zhao, Y.; Yang, S. Fine-Grained Air Pollution Inference at Large-Scale Region Level via 3D Spatiotemporal Attention Super-Resolution Model. *Atmosphere* **2025**, *16*, 166. https://doi.org/10.3390/atmos16020166

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Air pollution presents a serious hazard to human health and the environment for the global rise in industrialization and urbanization. The World Health Organization

(WHO) [1] estimates that nearly all of the world's population (99%) breathes air that is unhealthy and exceeds WHO air quality standards. Every year, 7 million premature deaths are attributed to the effects of household and ambient air pollution combined [2]. Air pollution usually shows regional characteristics, on the one hand due to inter-regional pollutant interactions caused by atmospheric circulation, and on the other hand due to a long history of irrational industrial layout in urban areas and focus on rough economic development. In particular, China has experienced increasing levels of $PM_{2.5}$ pollution in recent decades caused by rapid urbanization and industrialization [3,4]. Fine-grained $PM_{2.5}$ monitoring is crucial for understanding the formation and control of air pollution and their effects on human health. Furthermore, the "Chinese Government Work Report 2022" states that regional collaborative management monitoring for air pollution needs to be strengthened, and fine-grained pollutant monitoring is needed for proper urban management and air pollution source control [5].

The current research of air pollution inference mainly focuses on two scales: the macro national or regional level and micro city ground-level. For the macro national or regional level air pollution inference, satellite remote sensing techniques have been frequently used to estimate spatially continuous near-surface pollution concentrations for their capacity to track regional pollutant distributions [6–9]. Due to the limited data grained and the geographical proximity assumption, it is easy to introduce cumulative errors in statistical regression models and produce poor estimation results at the fine-grained level. For the micro city ground-level air pollution inference, monitoring air pollution with fine spatial granularity requires numerous devices to cover the target area, which brings a high cost of installation and maintenance [10–13]. The majority of the ground-level air pollution monitoring stations are distributed in urban centers and around industrial areas, and less in suburban and rural areas. This quantity distribution of air pollution monitoring stations is incredibly unbalanced, making it hard to obtain the unbiased estimation of air pollution at a large-scale region level.

In recent years, deep learning has made great progress in the super-resolution inference problem, which has motivated many applications in other fields, such as meteorology [14], climate [15], urban hotspots events prediction [16–19], and urban flow computing [11]. The video super-resolution model can solve the fine-grained air pollution inference problem to a certain extent, but they ignore the diffusion characteristics of air pollution [20]. Moreover, the correlation between frames in the video super-resolution problem is very high leading to feature redundancy, but there is no such problem in fine-grained air pollution inference [21].

Although much progress has been made by these methods, the existing macro-regional level or ground-level methods make air pollution inference by conducting interpolations from partial observations in the same spatial scale and fail to address the spatiotemporal correlations between cross-grained air pollution distribution. Furthermore, they still have limitations and there remain gaps when dealing with the issue of fine-grained air pollution inference (FAPI) at a large-scale region level, which can be embodied as follows:

(1) At the large-scale region level, the spatial and temporal distribution of pollutants in a region can be interpreted as one industrial area source in coarse-grained view. While in fine-grained view, it can be regarded as multiple plants emission forming multiple sources, indicating the pollutant dispersion heterogeneity in spatial and temporal distribution [22]. Furthermore, both coarse-grained and fine-grained distribution have strong local correlation due to pollution sources, wind direction, and topography [23]. Furthermore, the spatial local correlation between the distribution of coarse-grained and fine-grained pollution can vary significantly, whereas the spatial local correlation of coarse-grained pollution may decrease as a result of the more uniform distribution of pollution sources over a larger area. Due to the spatiotemporal variety distribution

and spatial local correlation of pollutant dispersion processes at different scales, it is more challenging to make fine-grained inferences about atmospheric pollutants.

(2) Atmospheric pollutants have a time-varying diffusion effect,and in the case of long time intervals (e.g., 1 day), pollutants are likely to experience significant changes, which makes it challenging to accurately infer the trend of pollutant diffusion and extract the characteristics of the short-term diffusion trend from the temporal distribution of atmospheric pollution [24,25].

(3) Pollutants emitted from one region can have long-lasting impacts on another region due to atmospheric diffusion and transport at a large scale [26]. Furthermore, due to the atmospheric circulation and various other factors, pollutants move through the atmosphere at a relatively slow rate of diffusion and deposition. Additionally, during the deposition process, pollutants can undergo transformations into different chemical species, leading to the persistence of certain pollutants in the environment [27]. Wind speeds and other meteorological conditions can facilitate the long-distance transport of atmospheric pollutants, allowing them to affect regions that are far removed from their original sources. Furthermore, it is essential to consider the long-term trends of pollutants over time when assessing their environmental impact.

Therefore, this study aims to propose a fine-grained air pollution inference framework to realize high-resolution air pollution estimation at large-scale region level. We convert the fine-grained air pollution inference as a single image super-resolution problem, and we propose a spatiotemporal super-resolution data-driven method to enhance the pollution map granularity. The proposed framework creates a mapping from the coarse-grained air pollutant regional distribution to get the fine-grained distribution, and the distribution of $PM_{2.5}$ pollutants in the Yangtze River Delta region on 31 December 2021 is shown in Figure 1 (the dataset details are shown in Table 1), where the coarse-grained pollutant distribution (8 km × 8 km) is shown on the left and the fine-grained pollutant distribution (1 km × 1 km) is shown on the right.

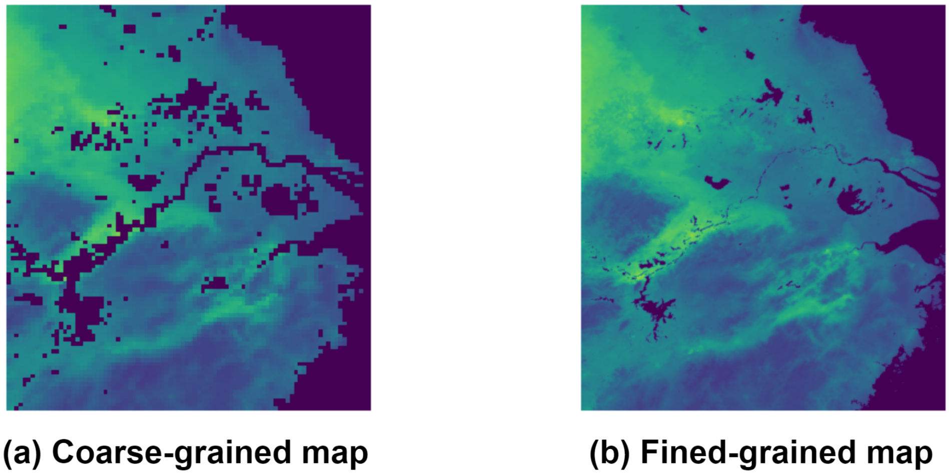


Figure 1. The distribution of $PM_{2.5}$ pollutants in the Yangtze River Delta region on 31 December 2021; (**a**) shows the coarse-grained distribution (8 km × 8 km) and (**b**) shows the fine-grained distribution (1 km × 1 km), where black represents the map boundary, which we set to -inf [28].

In this paper, a fine-grained air pollution inference model (AirSTFM) is proposed at the large-scale region level. The main contributions of this paper are as follows:

(1) A 3D-patch-wise self-attention convolutional module is designed to extract the spatiotemporal features of air pollution, which aggregates both spatial and temporal

information of coarse-grained air pollution and employs a sliding window to add spatial local features.

(2) A bidirectional optical flow feed-forward layer is designed to extract the short-term air pollution diffusion characteristics, which can learn the temporal correlation contaminant diffusion between closeness time intervals.

(3) A spatiotemporal super-resolution upsampling pretext task is constructed to model the long-term higher-level dispersion features mapping between the coarse-grained and fined-grained air pollution distribution.

(4) Comprehensive experiments are performed on the $PM_{2.5}$ pollution datatset of the Yangtze River Delta region, and the experiment results show that our proposed AirSTFM outperforms the state-of-the-art models.

Table 1. Details of dataset.

Dataset	
Air pollutant	$PM_{2.5}$
Time span	1 January 2013–31 December 2021
Time interval	1day
Coarse-grained size	100 × 90
Coarse-grained scale	8 km
Fine-grained size	800 × 720
Fine-grained scale	1 km
Upscaling factor	8
Latitude range	29°20′N–32°34′N
Longitude range	115°46′E–123°25′E

The rest of this paper is arranged as follows. In Section 2, we present the related work of the macro and micro air pollution inference model. We clarify the fine-grained air pollution inference problem and relevant definitions in Section 4. In Section 5, we present the detail components of the proposed AirSTFM model. The experiments and result analysis are presented in Section 6. Finally, we conclude the paper and discuss the future possible extending work in Section 7.

2. Related Work

2.1. Macro National/Regional Level Air Pollution Inference

For the macro national or regional level air pollution inference, satellite remote sensing techniques have been frequently used to estimate spatially continuous near-surface pollution concentrations for their capacity to track regional pollutant distributions. For instance, aerosol optical depth (AOD) is produced by a number of satellite sources, including Himawari-8 [6], the Visible Infrared Imaging Radiometer (VIIRS) [7], and Moderate Resolution Imaging Spectroradiometer (MODIS) [8]. The currently popular AOD products have a poor spatial resolution (between 3 and 10 km), and they exhibit significant estimation uncertainty on bright surfaces [9]. Additionally, the spatial and temporal distribution of $PM_{2.5}$ is incredibly complex since it is influenced by both natural and human influences. Furthermore, Jiang et al. [29] investigated the temporal and spatial aspects of air quality in China's Yellow River Basin economic zone. Yuan et al. [30] evaluated the features of air pollution changes in the Yangtze River Delta region during the COVID-19 outbreak. Furthermore, to analyze the air pollution of China's urban cluster regions, Deng et al. [31] assessed the geographical and temporal variability of $PM_{2.5}$ and the surface ozone. He et al. [32] used an enhanced gradient boosting decision tree to predict the regional distribution of $PM_{2.5}$ in China. Chen et al. [33] estimated the spatiotemporal distribution of PM_{10} pollution in

China using a random forest model. Wu et al. [34] integrated Kriging and inverse distance weighting (IDW) to realize adaptive interpolation of coarse-grained pollution maps. Blanchard et al. [35] interpolated spatiotemporally with weights calculated from inter-station pollutant correlations to generate daily air pollution concentration distributions at the 1–10 km scale. Wei et al. [36,37] suggested a tree-based ensemble spatial–temporal extra-tree (STET) model to generate high-quality, high-resolution $PM_{2.5}$ distributions for China from 2000 to 2020. Babaan et al. [38] used an ensemble hybrid spatial model to estimate ozone pollution in the Taiwan Province area. Yang et al. [39] proposed attention-based domain spatiotemporal meta-learning (ADST-ML) to adaptively extract the spatiotemporal dependence of $PM_{2.5}$ for a regional $PM_{2.5}$ prediction in Beijing. However, due to the limited data grained and the geographical proximity assumption, it is easy to introduce cumulative errors in statistical regression models and produce poor estimation results at the fine-grained level.

2.2. Micro City Ground-Level Air Pollution Inference

For the micro city ground-level air pollution inference, monitoring air pollution with fine spatial granularity requires numerous devices to cover the target area, which brings a high cost of installation and maintenance. In recent years, mobile sensor networks have been applied to obtain fine-grained sensing with fewer sensors. These sensors can be carried and powered by vehicles, e.g., taxi, tram, or bus, so that they can move around the city to sense the air quality at different locations [10,11]. Do et al. [12] used graph neural networks to infer fine-grained pollution distributions. Dun et al. [13] proposed a novel deep learning model by combining dynamic graph convolutional and multichannel spatiotemporal convolutional networks (DGC-MTCNs) for air quality prediction. Liu et al. [40] proposed spatial–temporal causal convolutional networks to make precise projections of the AQI in Shanghai, where spatial effects of several sources of air pollution and meteorological conditions were taken into account. Ma et al. [41] estimated the fine-grained distribution of $PM_{2.5}$ using a random forest method. Xu et al. [42] employed spatiotemporal graph convolutional networks to make road-level emission predictions. Zhang et al. [43] developed a hybrid deep learning framework that enables fine-grained air pollution estimation at the city-level. Hu et al. [44] recently combined feature recovery, feature extraction, and air quality inference into one model to infer the fine-grained distribution of air quality in Beijing. Hofman et al. [45] recently adopted bicycle mobile source sensors as a data source using a variational auto-encoder (AVGAE) and geographic random forest model (GRF) to increase the geographical monitoring resolution. Marjovi et al. [46] proposed three modeling approaches using large-scale mobile sensor networks to handle the dynamic coverage of mobile sensor networks to generate high spatial and temporal resolution maps of urban environmental pollutants. However, the majority of the ground-level air pollution monitoring stations are distributed in urban centers and around industrial areas, and less in suburban and rural areas. This quantity distribution of air pollution monitoring stations is incredibly unbalanced, making it hard to obtain the unbiased estimation of air pollution at a large-scale region level.

3. Study Area

The study area in this work is the Yangtze River Delta region, which is located in eastern China, comprising the provinces of Anhui, Jiangsu, and Zhejiang together with the municipality of Shanghai, as shown in Figure 2. Furthermore, it is one of the most developed and prosperous regions in China. The region is one of China's key economic engines, with abundant resources and a unique geographical location. The core cities of the

Yangtze River Delta include Shanghai, Nanjing, Hefei, and Hangzhou. This region is flat and characterized by extensive farmland, waterways and urban agglomerations.

The Yangtze River Delta region is influenced by a monsoon climate with four distinct seasons. Summers are warm and humid, and winters are relatively cold. The Yangtze River Delta region has well-developed port facilities and many important ports on the Yangtze River, including the Port of Shanghai, which is one of the largest container ports in the world. These ports play an important role in regional and global trade and provide solid support for China's economic growth.

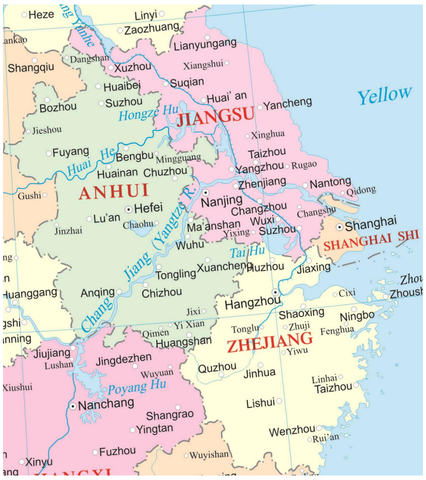


Figure 2. The map identifies the study area of the Yangtze River Delta region, where the distribution of $PM_{2.5}$ is concentrated [47].

4. Preliminary

4.1. Coarse/Fine-Grained Pollution Map

The coarse-grained pollution map indicates the pollution data granularity in the observation sequence. As shown in Figure 3a, the resolution of the coarse-grained pollution is 5 × 5 (2 km in spatial resolution). The fine-grained pollution map is the target data granularity to predict, where the fine-grained resolution is 10 × 10 (1 km in spatial resolution) as shown in Figure 3b. For a target region area where the air pollution map uniformly divides into $H \times W$ grid maps, $X_t \in R^{H \times W}$ denotes the pollution contaminants at a certain time t. Figure 3 shows the coarse-grained and fine-grained pollution maps with an upsampling factor of r = 2, and the coarse-grained pollution map is obtained by integrating the average of r^2 nearby grid contaminants in a fine-grained pollution map. In our work, the resolution of the coarse-grained pollution is set to 8 km in spatial resolution with an upsampling factor of r = 8, which is shown in Figure 1.

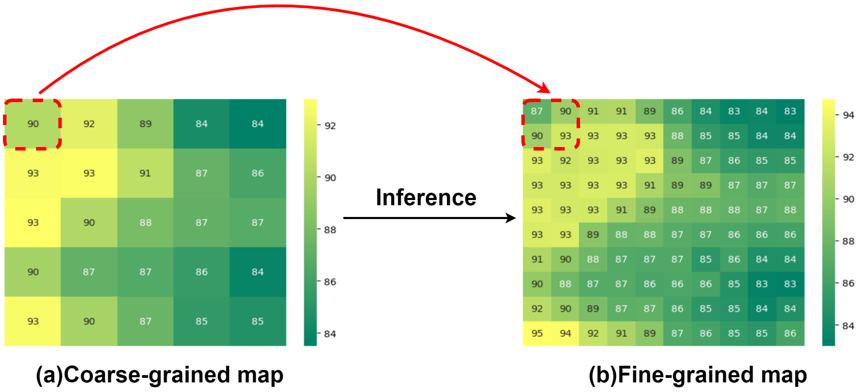


Figure 3. The illustration of the coarse-grained and fine-grained pollution map.

4.2. Problem Formulation

Therefore, the fine-grained air pollution inference problem (FAPI) can be defined as follows, given upsampling factor r and the coarse-grained spatiotemporal air pollution map $X^C = \{X_1^c, X_2^c, X_3^c \dots X_t^c\}$,$X_t^c \in R^{H\times W}$, to make an inference of the corresponding fine-grained map $X^F = \{X_1^F, X_2^F, X_3^F \dots X_t^F\}$, $X_t^F \in R^{rH\times rW}$.

$$\hat{X}^F = f(X^C;\theta) = arg \max_{X^F} p(X^F|X^C) \tag{1}$$

5. Methodology

In this section, we propose a 3D spatiotemporal attention super-resolution model (AirSTFM) for fine-grained air pollution inference at a large-scale region level, as shown in Figure 4. Firstly, the $PM_{2.5}$ pollution map in the target region at each time interval is converted into a fined-grained map X^F and coarse-grained map X^C, respectively. A spatiotemporal super-resolution pretext task is constructed to extract long-term spatiotemporal dispersion characteristics of more coarse-grained pollution maps. Furthermore, a 3D-patch-wise self-attention convolutional module is designed to extract the spatiotemporal features of air pollution, which aggregates both the spatial and temporal information of coarse-grained air pollution and employs a sliding window to add spatial local features. Then, a bidirectional optical flow feed-forward layer is used to extract the short-term air pollution diffusion characteristics, which can learn the temporal correlation contaminant diffusion between closeness time intervals. Finally, we construct a upsampling reconstruction module to model the higher-level features mapping between the coarse-grained and fined-grained air pollution distribution.

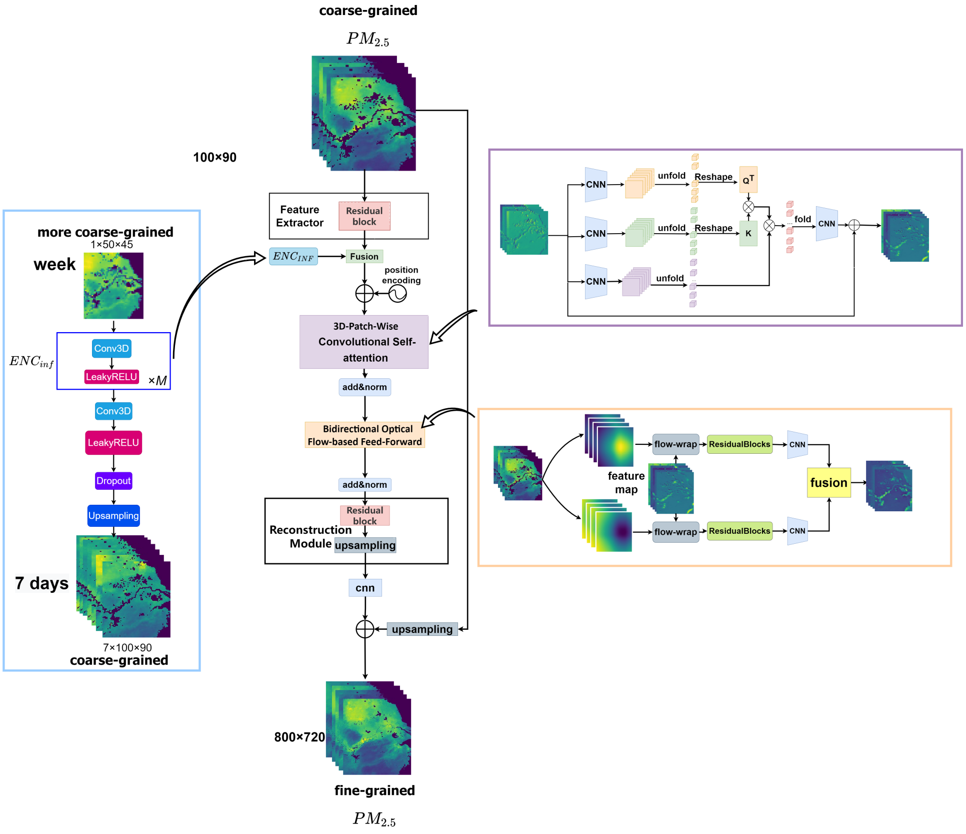


Figure 4. The framework of AirSTFM, which contains three modules, a 3D-patch-wise convolutional self-attention module,bidirectional optical flow-based feed-forward module, and a spatial–temporal super-resolution inference network, for extracting high-level contaminant semantic information using a pretext task.

5.1. Spatiotemporal Super-Resolution Pretext Task

Considering the dispersion of pollutants is often long lasting, but capturing the dispersion characteristics of pollutants over long periods of time is often very difficult. Here, we construct a spatiotemporal super-resolution pretext task to pre-train the inference network. Our aim is to use a mapping that is more coarse-grained in both time and space to learn higher-level features at both spatial and temporal levels. Specifically, we take the coarse-grained map $X_C^{T\times H\times W}$ and continue upsampling to obtain a more coarse-grained $X_{MC}^{\frac{T}{S_t}\times\frac{H}{S_h}\times\frac{W}{S_w}}$ contaminant distribution map. Then, we can construct a spatiotemporal super-resolution pretext task to map from $X_{MC}^{\frac{T}{S_t}\times\frac{H}{S_h}\times\frac{W}{S_w}}$ to $X_C^{T\times H\times W}$, which can help us explore the long time pollutant dispersion characteristics.

The super-resolution pretext task is shown in Figure 5. We use multiple 3D convolutional and relu layers to capture spatiotemporal features, which is followed by upsampling using the same method, and the final loss function is adopted as RMSE. By inferring a more coarse-grained to coarse-grained mapping, higher-level feature semantic information can be learned, especially the distribution in time, and we can better extract the spatiotemporal diffusion characteristics of pollutants.

$$\mathcal{L}_{inf} = \frac{1}{T}\sum_{t=1}^{T}\left\|\mathbf{X}_t^C - Y_{ENC}\left(\mathbf{X}_t^{MC};\theta\right)\right\|^2 \tag{2}$$

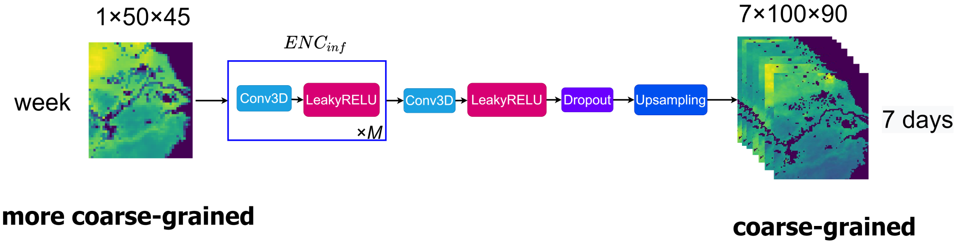


Figure 5. Illustration of super-resolution pretext task.

5.2. 3D-Patch-Wise Self-Attention Convolutional Module

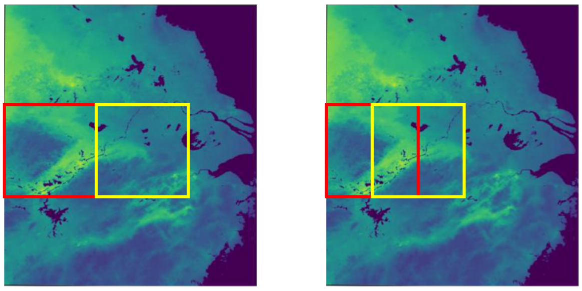
The vision transformer (VIT) [48] model introduces the self-attention mechanism to the image processing aspect, and the method usually cuts the image into a non-overlapping patch, which will lose some original spatial structure and local features. In the FAPI task, texture and details need to be reconstructed, as shown in Figure 6. The VIT model cuts the images into non-overlapping blocks, and the pollutants have certain spatial characteristics of diffusion which will inevitably disrupt some spatial structure of the original image, while our self-attention module overlaps the chunks, which not only retains the spatial structure, but also increases certain local correlation. If the space is cut directly using the VIT model, the spatial features of atmospheric pollutant diffusion will be lost, for which we propose a 3D-patch-wise convolutional self-attention module.

As shown in Figure 7, we put the input air pollutant features X into three independent convolutional networks to extract spatial information. Furthermore, subsequently, we use sliding local 3D patches with step stride=1 and T*Wp*Hp patch sizes from the feature maps for the unfold operation, which can better extract local correlations. Then, we reconstruct the obtained 3D patch into a one-dimensional feature vector to obtain queries subspace (Q), key subspace (K), and calculate the similarity matrix using dot product, and aggregated with value subspace (V) into a feature map. Here, it is important to note that because a 3D patch sliding window is used, the obtained feature map contains the spatiotemporal diffusion features of the $PM_{2.5}$ pollutants.

$$\begin{aligned} Q &= f_{unfold}(W_Q \times X) \\ K &= f_{unfold}(W_K \times X) \\ V &= f_{unfold}(W_V \times X) \end{aligned} \tag{3}$$

$$\begin{aligned} f_{attention}(\mathcal{X}) = \phi(\mathcal{X} + \sum_{i=1}^{h} \mathcal{W}_o^i * f_{fold}(\underbrace{f_{unfold}\left(\mathcal{W}_V^i * \mathcal{X}\right)}_{\mathcal{V}} \\ \sigma_1(\underbrace{f_{unfold}\left(\mathcal{W}_K^i * \mathcal{X}\right)^T}_{\mathcal{K}} \underbrace{f_{unfold}\left(\mathcal{W}_Q^i * \mathcal{X}\right)}_{\mathcal{Q}}))) \end{aligned} \tag{4}$$

Finally, as shown in Figure 8, we use the fold operation to reassemble the 3D patch into a feature map with the same dimensions as the original feature map. The feature map with the spatial and temporal diffusion characteristics of pollutants can be obtained by performing the unfold operation on the convolved feature map and then performing the attention mechanism. Considering the sliding window will overlap, we average the overlap between the patches.



(a)VIT

(b)Ours

Figure 6. (**a**) The VIT divides images into blocks without overlap while pollutants affect spatial structure. (**b**) Our self-attention models overlap the chunks, which retains the structure and enhances local correlation.

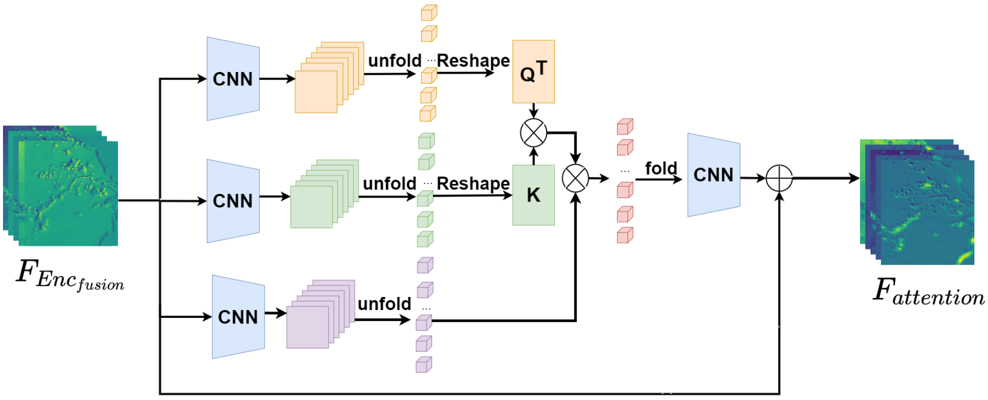


Figure 7. Illustration of 3D-patch-wise self-attention convolutional module, where the input $F_{Enc_{fusion}}$ are fusion features generated by the spatiotemporal super-resolution pretext task. The unfold operation uses a sliding window to decompose and combine feature maps into patches, and the fold operation is used to recombine patches into feature maps $F_{attention}$.

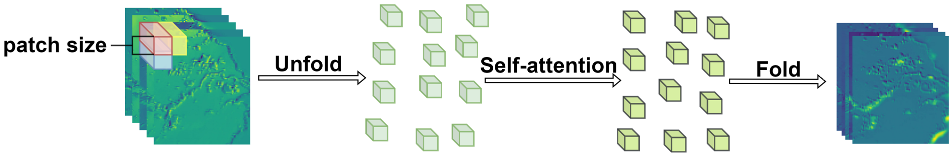


Figure 8. Unfolding the convolved spatial–temporal diffusion feature maps.

5.3. Bidirectional Optical Flow-Based Feed-Forward Layer

The conventional transformer uses a fully connected layer in the feed-forward layer, which neglects the temporal relationships between tokens and cannot achieve alignment between images. Inspired by BasicVSR [49], we propose a feed-forward layer based on bidirectional optical flow to model the local correlation of the diffusion time of pollutants, which ensures spatial alignment of the diffusion motion of $PM_{2.5}$ pollutants, as shown in Figure 9.

$$\overleftarrow{\boldsymbol{O}}_t = \begin{cases} s(\boldsymbol{V}_1, \boldsymbol{V}_1), & \text{if } t = 1, \\ s(\boldsymbol{V}_{t-1}, \boldsymbol{V}_t), & \text{if } t \in (1, T] \end{cases} \tag{5}$$

$$\overrightarrow{\boldsymbol{O}}_t = \begin{cases} s(\boldsymbol{V}_{t+1}, \boldsymbol{V}_t), & \text{if } t \in [1, T) \\ s(\boldsymbol{V}_T, \boldsymbol{V}_T), & \text{if } t = T \end{cases} \tag{6}$$

$$\overleftarrow{\mathcal{X}} = \omega(\mathcal{X}, \overleftarrow{\mathcal{O}}), \overrightarrow{\mathcal{X}} = \omega(\mathcal{X}, \overrightarrow{\mathcal{O}}) \tag{7}$$

where s represents our bidirectional optical flow estimation algorithm [50], V_i represents the atmospheric pollution distribution at frame i in the current time window, and w represents the wrapping function.

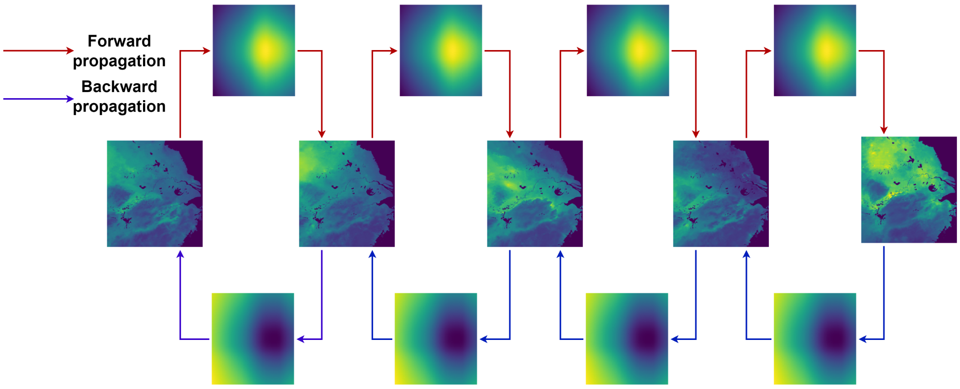


Figure 9. Optical flow estimation of the pollutant distribution at adjacent time points.

Then, we use convolutional pyramids to estimate the optical flow between frames $\overrightarrow{\boldsymbol{O}}$ and $\overleftarrow{\boldsymbol{O}}$. The output X from the spatiotemporal convolutional self-attention layer is aggregated using the wrapping method to obtain $\overrightarrow{\mathcal{X}}$ and $\overleftarrow{\mathcal{X}}$ forward propagation features and backward propagation features from the optical flow information and data.

$$\begin{aligned} f_{output}(\mathcal{X}) = \phi(f_{attention}(\mathcal{X}) + \rho(ResidualConv(\mathcal{V}, \overleftarrow{\mathcal{X}}) \\ + ResidualConv(\mathcal{V}, \overrightarrow{\mathcal{X}}))) \end{aligned} \tag{8}$$

Furthermore, as shown in Figure 10, in order to learn the relationship between frames, we improve the fully connected feed forward layer by improving it to convolutional forward and backward propagation, where $\rho(-)$ is the fusion operation. By the bidirectional propagation of the optical flow, thus being able to learn the temporal correlation contaminant diffusion between time intervals.

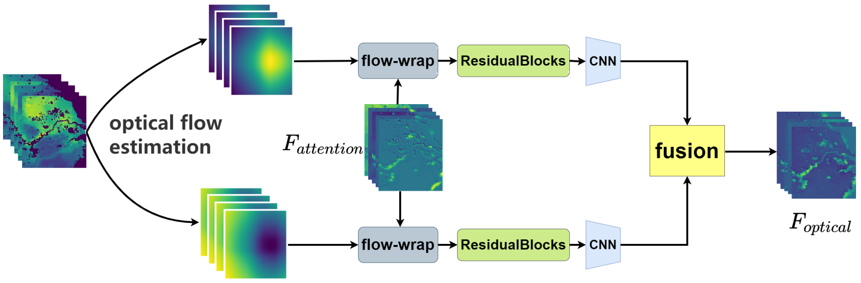


Figure 10. The optical flow map is combined with the feature map $F_{attention}$ from the attention mechanism. Furthermore, a two-way propagation network then generates separate feature maps, which are fused into $F_{optical}$.

5.4. Feature Reconstruction

After obtaining the features for optical flow fusion, we use the residual block to extract the fused features $F_{optical}$ as shown in Figure 11. For the input contamination distribution feature tensor, we apply a convolutional layer to process it. The convolutional layer includes 64 input channels and 64 output channels with a convolutional kernel size of 3 × 3, a step size of 1, and a padding of 1. Subsequently, we use a LeakyReLU activation function to retain the extracted feature information. Next, we perform multiple upsampling operations to convert the features from coarse-grained to fine-grained level. Each upsampling section consists of the following steps: (1) using a convolutional layer with 64 input channels, 256 output channels, a convolutional kernel size of 3 × 3, a step size of 1, and a padding of 1; (2) performing a Pixel-Shuffle operation on the outputs of the above convolutional layer to increase the spatial dimensions while decreasing the channel dimensions in order to achieve the effect of upsampling; and (3) the output of the above operation the Leaky-ReLU activation function is again applied to ensure the continuity and expressiveness of the features. Repeating the upsampling operation three times helps to gradually transform the low-resolution contamination distribution features into high-resolution features while retaining important information about the details. Furthermore, the final loss function is shown in Equation (9).

$$\mathcal{L}_{final} = \frac{1}{T}\sum_{t=1}^{T}\left\|\hat{X}_t^F(\theta) - X_t^F\right\|^2 \tag{9}$$

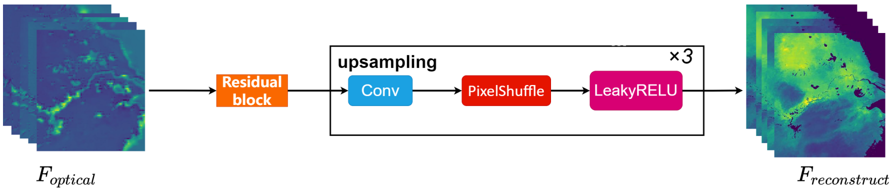


Figure 11. The feature maps obtained from the optical flow were residualized and upsampled to obtain fine-grained spatial and temporal distributions of the feature contamination.

6. Experiments and Results

6.1. Experiment Settings

Our study area covers different geographical features of the Yangtze River Delta region. We obtained data from 1 January 2013 to 31 December 2021 from the seamless spatial distribution dataset of $PM_{2.5}$ in the Yangtze River Delta region, which was provided by several sources, including [36,37]. The details of the dataset are shown in Table 1. Furthermore, we divide these data into a training set, validation set, and test set in 2:1:1 for better analysis and modeling. With these data, we can study the air quality trends in the Yangtze River Delta region and provide an important reference for improving the environmental quality and living conditions in the region.

Our model was deployed in pytorch1.10.2 and CUDA11.3 on RTX3090. Our model used the ADAM optimizer with parameters set to 0.9 and 0.99. Furthermore, the learning rate was set to 2×10^{-4}. The number of transformer modules was 1, the number of feature extraction blocks was 5, the number of feature reconstruction blocks was 30, and the length of the input time window was 5, the number of transformer modules was 1, the number of feature extraction blocks was 5, and the number of feature reconstruction blocks was 30. Batch_size = 2, and its base_channels were 32, the sliding window was $5 \times 10 \times 10$, and the number of M was set to 3. For a fair comparison, our model experiments are compared in the next five time points.

6.2. Baselines

MEAN: We distributed the pollutant data directly and uniformly to the fine granularity without feature extraction.

HA: Similar to the above method, but we added the time dimension so that it was averaged both spatially and temporally.

Urban-FM [51]: The authors utilized a feature extraction module and a new upsampling module for distribution to generate fine-grained flow distributions from coarse-grained inputs and utilize a generic fusion sub-network to further improve performance by taking into account the effects of different external factors. The performance was good on traffic flow and pedestrian flow datasets.

EDVR [52]**:** EDVR proposes solutions to two common difficulties in the video overdetermination problem: content alignment and feature fusion. For the alignment problem, a pyramidal cascaded deformable convolutional alignment network (PCD) was proposed, based on a deformable convolutional DCN, whose multi-level cascade setting produced a structure from rough to accurate estimation. For the feature fusion problem, the spatiotemporal attention fusion SR network (TSA) was proposed, which was based on an attention mechanism that focuses on important information and ignores useless or erroneous information by assigning different weights to the information contained in different frames and different spatial structures due to their different importance to image reconstruction.

basicVSR [49]**:** By designing a bidirectional loop structure of propagation, feature-wise alignment based on optical flow, and using some existing fusion and upsampling methods, the authors came up with a simple and lightweight video super segmentation method that outperforms existing VSR structures in terms of speed and reconstruction performance.

IconVSR [49]**:** Based on basicVSR, the authors proposed a module containing two new extensions to improve aggregation and propagation components. The first module is called information-fill, a mechanism that uses an additional module to extract features from sparsely selected frames (key frames) and then inserts these features into the main network for feature refinement. The second extension is a coupled propagation scheme that facilitates the exchange of information in the forward and backward propagation branches. These two modules not only reduce the accumulation of errors due to occlusions and image boundaries during propagation, but also allow the propagation to access complete information in a sequence for generating high-quality features.

basicVSR++ [53]**:** The authors improved the Propagation and Alignment sections based on the BasicVSR above. Specifically, Grid-Propagation was used in the feature propagation (Propagation) to repeatedly correct the alignment accuracy, and in the alignment (Alignment) section, a cross-grid propagation mechanism with second-orderMarkov property and a deformable convolutional alignment module with optical flow guidance were proposed.

MANA [54]**:** The model was designed with a cross-frame non-localattention mechanism that allows VSR to be more robust to large movements in the video without frame alignment. To obtain information beyond adjacent frames, a new memory-augmented attention module was additionally designed to memorize general video details during the training of SR.

6.3. Evaluation Metrics

To fully evaluate the performance of our proposed network, we used Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) to measure the prediction performance.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{X}_i)^2} \tag{10}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - \hat{X}_i\right| \tag{11}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{X_i - \hat{X}_i}{X_i}\right| \tag{12}$$

6.4. Results Analysis

We first evaluated our own model against the baseline on different scaled training datasets. The comparison is shown in Table 2. Note here that we omitted the variance in the table because the standard deviations of our test results are in [0, 5%].

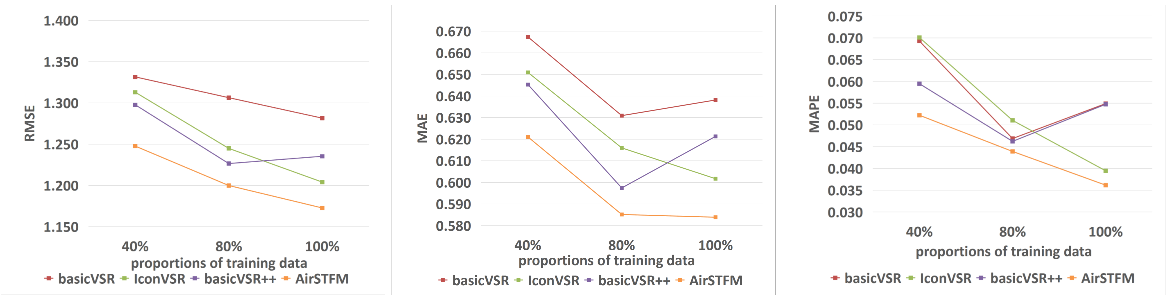
Table 2 shows the RMSE of the validation set when trained with a 100% partitioned dataset. When compared with other baselines, the AirSTFM converged faster and smoother, and the loss values were kept low in the early stage, which indicates that our pretext task successfully extracted the pollutant dispersion features and fused them with the model.

Table 2. Testing data performance with different training proportions.

	100%			**80%**			**40%**		
Methods	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
Mean	24.710	16.227	1.025	24.710	16.227	1.025	24.710	16.227	1.025
HA	5.039	1.942	0.534	6.760	2.049	0.592	6.479	1.999	0.546
Urban-FM [51]	4.808	2.021	0.573	5.970	3.536	0.237	9.520	5.877	0.438
EDVR [52]	1.386	0.737	0.061	1.403	0.741	0.064	1.991	1.032	0.126
basicVSR [49]	1.281	0.638	0.055	1.306	0.631	0.047	1.331	0.667	0.069
IconVSR [49]	*1.204*	*0.602*	*0.039*	1.245	0.616	0.051	1.313	0.651	0.070
basicVSR++[53]	1.235	0.621	0.055	*1.226*	*0.597*	*0.046*	*1.298*	*0.645*	*0.059*
MANA [54]	1.711	1.191	0.111	1.896	1.599	0.184	4.589	1.885	0.428
AirSTFM Δ	1.173 +2.60%	0.584 +3.05%	0.036 +6.36%	1.200 +2.17%	0.585 +2.06%	0.044 +5.01%	1.248 +3.86%	0.621 +3.76%	0.052 +12.18%

Furthermore, we divided the training set by 100%, 80%, and 40% to evaluate the effect of our model on datasets of different sizes in Table 2 and Figure 12. It can be seen that our model outperformed the other methods in all these divisions. In the 100% division, our model outperformed the second best model in RMSE, MAE, and MAPE by 2.6%, 3.05%, and 6.36%, and in the 40% division, our model also outperformed the second best model in RMSE, MAE, and MAPE by 3.86%, 3.76%, and 12.18%, which demonstrates the applicability of our model with different data sizes.

It is shown by the above results that our model achieved a very large advantage at different data scales. This is exactly in line with our motivation that our model makes better use of the air pollution dispersion characteristics compared to other models in the FAPI problem. In the Figure 13, to better compare the models, we picked the $PM_{2.5}$ distribution on a certain day in January of a certain year, where we can see that the $PM_{2.5}$ pollution was serious in the north due to winter heating and other reasons, while there were many heavy chemical enterprises distributed in the middle reaches of the Yangtze River basin and the regional pollutant distribution was all more serious. In the Mean method, the effect was poor, which shows the importance of capturing the spatial and temporal dispersion characteristics of air pollution. In the comparison with Urban-FM, our model outperformed it in terms of effect. As we can see in Figure 13, although Urban-FM and HA's inferred loss values are similar, they present very different effects: HA is seriously missing in details, while Urban-FM restores the details in place, but the effect on the boundary is poor. This is because Urban-FM mainly deals with a fine-grained urban flow inference (FUFI) problem, which has special structural constraints and is more suitable for traffic flow and pedestrian flow problems, which is not compatible with atmospheric pollutant dispersion characteristics. Most of the other state-of-the-art models tend to focus on the video overdetermination problem, which uses a very large correlation between frames in the dataset, often resulting in the problem of feature redundancy, while in the FAPI problem, the time interval is 1 day, and the temporal features are difficult to extract, resulting in the video overdetermination model being relatively unsuitable in the FAPI problem.



(**a**) Results on RMSE (**b**) Results on MAE (**c**) Results on MAPE

Figure 12. Results compared to baseline methods for different training set sizes.

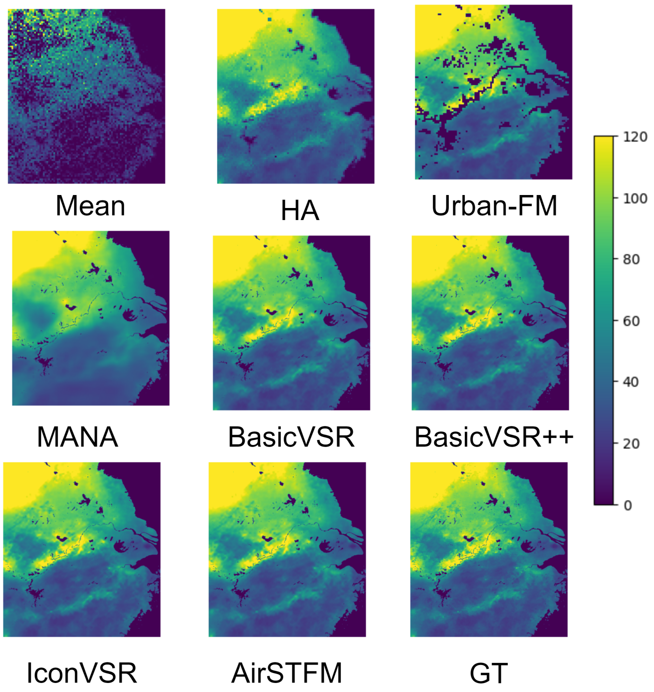


Figure 13. Visualization of the comparison of our method with each baseline method for the inferred $PM_{2.5}$ distribution, where GT stands for the ground truth.

Figure 14 shows the comparison of the number of parameters and the loss in the 100% division of the dataset between our model and some video hyperdivision models. The table shows the comparison between the AirSTFM and baseline methods in terms of training time, and in comparison with IconVSR, it achieved the second highest inference result, but the training time was twice as long as ours. However, since the MANA model focuses on solving large motion models and does not use an alignment mechanism, it is less effective on the FAPI problem.

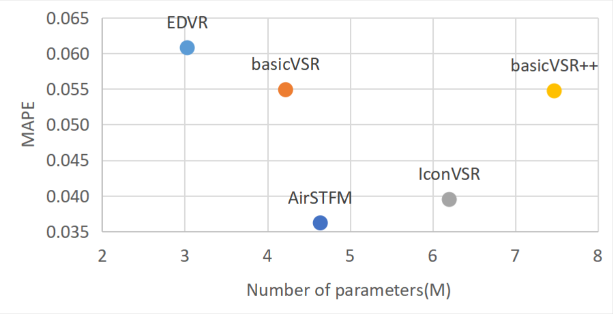


Figure 14. Model parameters' comparison.

Furthermore, Figure 15 and Table 3 shows the convergence speed comparison of different models, where the RMSE of the validation set was trained with a 100% partitioned dataset. When compared with other baselines, the AirSTFM converged faster and smoother, and the loss values were kept low in the early stage, which indicates that our pretext task successfully extracted the pollutant dispersion features and fused them with the model.

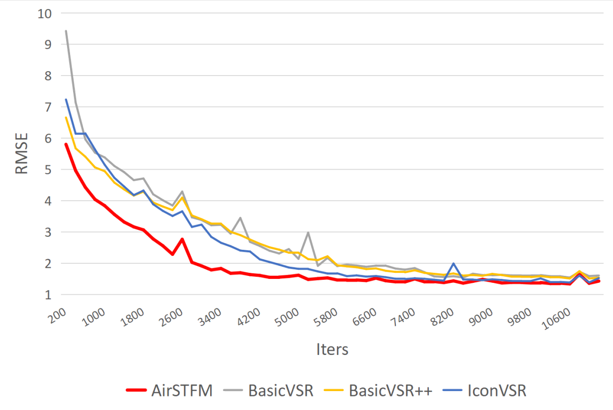


Figure 15. Convergence speed comparison of different models.

Table 3. Model parameters and training time comparison.

Models	Params (M)	Training Time	RMSE	MAE	MAPE
Urban FM	6.42	12 h 52 min	4.808	2.021	0.573
EDVR	3.03	10 h 41 min	1.386	0.737	0.061
basicVSR	4.22	30 h 6 min	1.281	0.638	0.055
IconVSR	6.2	43 h 21 min	1.204	0.602	0.039
basicVSR++	7.47	35 h 50 min	1.235	0.621	0.055
MANA	82.03	8 h 50 min	1.711	1.191	0.111
AirSTFM	4.64	20 h 6 min	**1.173**	**0.584**	**0.036**

Figure 16 shows the difference between the speculative and real values, where the higher brightness represents a larger error. It can be seen that most of them are distributed in the middle and lower reaches of the Yangtze River as well as coastal areas. This is due to the fact that the cities of An'qing and Tongling in the middle and lower reaches of the

Yangtze River, as heavy industrial cities, produce more serious pollution and rapid changes in air pollutant emissions, which makes the extrapolation more difficult and thus leads to larger errors.

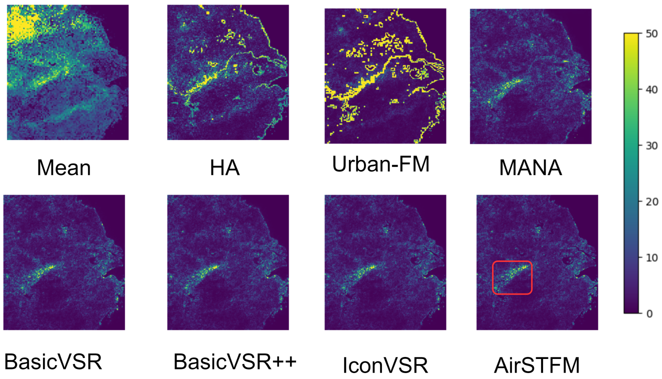


Figure 16. Visualization of the difference between inferred and actual values.

6.5. Ablation Experiments

In order to analyze the contribution of each module of our model, we tested the modules individually and in two-by-two combinations, under a 100% training set. It should be noted here that attention unchecked represents the replacement of our 3D-patch-wise convolutional self-attention with the attention mechanism of the original transformer, and optical flow unchecked represents the replacement of the optical flow module with the fully connected layer of the original transformer.

From Table 4, we can find that any combination of two or more modules works better than one module alone, and the combination of three works best. This proves the effectiveness of our component combination. When using separate modules, it can be seen that the attention mechanism works best, and when using a two-by-two combination of modules, it is seen that the module containing the attention module outperforms the module without the attention module. When using the optical flow module alone, the experiments are less effective; this is because when using the optical flow model alone, the estimated optical flow is wrappedwith the original feature maps extracted from the attention module and not fused with the features containing spatiotemporal variation, which is less effective. When combined with the 3D patch attention module, the effect improves significantly.

Table 4. Ablation studies.

3D-Patch-Wise Self-Attention Module	Bidirectional Optical Feed-Forward Layer	Spatial–Temporal Inference Network	RMSE	MAE	MAPE
√			1.316	0.642	0.048
	√		1.387	0.768	0.062
		√	1.369	0.802	0.060
√	√		1.213	0.596	0.044
√		√	1.209	0.599	0.042
	√	√	1.230	0.609	0.040
√	√	√	1.173	0.584	0.036

Table 5 shows a comparison of the fusion of the features extracted by the pretext task using different methods at different locations in the network structure (“before” represents the position after placing feature fusion in the feature extraction module and before position encoding, “after” represents the position after placing feature fusion in the feature reconstruction and before upsampling), with the best results using the add method and before the network structure. This is because the goal of our pretext task is to infer the spatiotemporal distribution of pollution from a more coarse-grained pollution distribution of 1 × 50 × 45 to 7 × 100 × 90 to infer the pollutant-specific spatiotemporal dispersion characteristics from the regional level, which requires early fusion of the features and fusion with the features at a high level.

Table 5. Different fusion strategies effects.

	RMSE	MAE	MAPE
AirSTFM (concat fusion "after")	1.248	0.625	0.056
AirSTFM (concat fusion "before")	1.221	0.61	0.05
AirSTFM (add fusion "after")	1.225	0.602	0.046
AirSTFM (add fusion "before")	1.173	0.584	0.036

6.6. Hyperparameters Analysis

In this section, we analyze some hyperparameters of the model, and each time we change a hyperparameter, we set the other parameters to their default values. The hyperparameter experiment results are shown in Tables 6–9.

Table 6. Different number of feature extract blocks.

Extract Blocks Number	RMSE	MAE	MAPE
3	1.321	0.879	0.054
5	1.173	0.584	0.036
10	1.294	0.868	0.043

Table 7. Different number of feature reconstruct blocks.

Reconstruct Blocks Number	RMSE	MAE	MAPE
10	1.267	0.712	0.046
30	1.173	0.584	0.036
50	1.328	0.887	0.052

Table 8. Different number of sequence lengths.

Sequence Length	RMSE	MAE	MAPE
3	1.298	0.859	0.044
5	1.173	0.584	0.036
7	1.297	0.861	0.043

Table 9. Different number of channels.

Channels	RMSE	MAE	MAPE
16	1.626	0.960	0.085
32	1.173	0.584	0.036
64	1.255	0.846	0.036

In Tables 6 and 7, we analyze the number of feature extraction and feature reconstruction, and we can observe from the table that the fine-grained inference works best when the number of extraction blocks is 5 and the number of reconstruction blocks is 30.

Furthermore, in Table 8, we can find that the performance increases when we change the sequence length from 3 to 5, but decreases when it is further increased to 7. It is possible that the excessive sequence length contains too much redundant time information, which introduces additional noise to the inference. From Table 9, we see that the best results are obtained when the number of channels = 32. When the number of channels was 64, the results became worse, which was due to the fact that increasing the number of channels may lead to over-fitting.

In order to analyze the different overlap rate effect in Figure 6, from Table 10 we can find that the higher the overlap rate between two adjacent chunks, the better the prediction performance. The overlap rate parameter is set to 90% in this work, which achieved the beat prediction performance.

Table 10. Different overlap rate effects.

Overlap Rate	RMSE	MAE	MAPE
90%	1.173	0.584	0.036
70%	1.207	0.591	0.041
50%	1.219	0.594	0.043
0	1.216	0.596	0.041

7. Conclusions

In this paper, we propose a 3D spatiotemporal attention super-resolution model (AirSTFM) for fine-grained air pollution inference at a large-scale region level. Different from the existing macro-regional level or ground-level methods which make air pollution inference in the same spatial scale, our AirSTFM employs a 3D-patch-wise self-attention convolutional module to extract the spatiotemporal features of coarse-grained air pollution. Moreover, a bidirectional optical flow feed-forward layer is designed to extract the short-term air pollution diffusion characteristics, which can learn the temporal correlation contaminant diffusion between closeness time intervals. Finally, we construct a spatiotemporal super-resolution upsampling pretext task to model the higher-level dispersion features mapping between the coarse-grained and fined-grained air pollution distribution. The proposed AirSTFM model is evaluated on the $PM_{2.5}$ pollution dataset of the Yangtze River Delta region, our model outperforms the second best model in RMSE, MAE, and MAPE by 2.6%, 3.05%, and 6.36% in the 100% division, and our model also outperforms the second best model in RMSE, MAE, and MAPE by 3.86%, 3.76%, and 12.18% in the 40% division, which demonstrates the applicability of our model at different data sizes. Furthermore, the comprehensive experiment results show that our proposed AirSTFM outperforms the state-of-the-art models.

In the future, we will extend our proposed model to address missing patterns in spatiotemporal air pollution data by introducing diffusion models framework. Moreover, we will apply the proposed model to other areas of smart city regulation, such as urban greenhouse gases prediction, urban precipitation forecasting, etc.

Author Contributions: Conceptualization, C.L.; methodology, C.L. and S.T.; software, C.L. and K.P.; validation, J.L.; formal analysis, K.P. and C.L.; investigation, C.L. and K.P.; resources, Y.Z.; data curation, K.P. and Z.X.; writing—original draft preparation, C.L.; writing—review and editing, Z.X. and Y.Z.; visualization, J.L. and K.P.; supervision, Z.X. and S.Y.; project administration, Z.X.; funding acquisition, Z.X. and S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Open Research Fund of Key Laboratory for Vehicle Emission Control and Simulation of the Ministry of Ecology and Environment, Chinese Research Academy of Environmental Sciences (VECS2024K03), and the Fundamental Research Funds for the Central Public-interest Scientific Institution (2024YSKY-03), National Natural Science Foundation of China (62033012), Open Project Program of Anhui Provincial Key Laboratory of Multimodal

Cognitive Computation (Anhui University) under grant MMC202303, Key Project of Anhui Provincial Department of Education under grant 2024AH050407, General Teaching and Research Project of Quality Engineering in Anhui Province under grant 2022jyxm423, and the General Project of Anhui Provincial Human Resources and Social Security Department under grant gxgnfx2022213.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study will be available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. World Health Organization. Air Pollution. Available online: https://www.who.int/health-topics/air-pollution (accessed on 22 January 2025).
2. World Health Organization. Ambient (Outdoor) Air Pollution 2024. Available online: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (accessed on 22 January 2025).
3. He, K.; Huo, H.; Zhang, Q. Urban Air Pollution in China: Current Status, Characteristics, and Progress. *Annu. Rev. Energy Environ.* **2002**, *27*, 397–431. https://doi.org/10.1146/annurev.energy.27.122001.083421.
4. Sun, L.; Wei, J.; Duan, D.; Guo, Y.; Yang, D.; Jia, C.; Mi, X. Impact of Land-Use and Land-Cover Change on urban air quality in representative cities of China. *J. Atmos. Sol.-Terr. Phys.* **2016**, *142*, 43–54. https://doi.org/https://doi.org/10.1016/j.jastp.2016.02.022.
5. Li, K. Report on the Work of the Government. 2023. Available online: https://www.google.com.hk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjX8vuAiaCLAxUhMDQIHSlfOzcQFnoECBcQAQ&url=https%3A%2F%2Fnpcobserver.com%2Fwp-content%2Fuploads%2F2023%2F03%2F2023-Government-Work-Report.pdf&usg=AOvVaw1p5K_ySIEgHiiraBUMWMw9&opi=89978449 (accessed on 22 January 2025)
6. Zhang, T.; Zang, L.; Wan, Y.; Wang, W.; Zhang, Y. Ground-level $PM_{2.5}$ estimation over urban agglomerations in China with high spatiotemporal resolution based on Himawari-8. *Sci. Total Environ.* **2019**, *676*, 535–544. https://doi.org/https://doi.org/10.1016/j.scitotenv.2019.04.299.
7. Yao, F.; Wu, J.; Li, W.; Peng, J. A spatially structured adaptive two-stage model for retrieving ground-level $PM_{2.5}$ concentrations from VIIRS AOD in China. *ISPRS J. Photogramm. Remote Sens.* **2019**, *151*, 263–276. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2019.03.011.
8. Ma, Z.; Hu, X.; Huang, L.; Bi, J.; Liu, Y. Estimating ground-level $PM_{2.5}$ in China using satellite remote sensing. *Environ. Sci. Technol.* **2014**, *48*, 7436–7444.
9. Wei, J.; Li, Z.; Guo, J.; Sun, L.; Huang, W.; Xue, W.; Fan, T.; Cribb, M. Satellite-derived 1-km-resolution PM1 concentrations from 2014 to 2018 across China. *Environ. Sci. Technol.* **2019**, *53*, 13265–13274.
10. Hasenfratz, D.; Saukh, O.; Walser, C.; Hueglin, C.; Fierz, M.; Arn, T.; Beutel, J.; Thiele, L. Deriving high-resolution urban air pollution maps using mobile sensor nodes. *Pervasive Mob. Comput.* **2015**, *16*, 268–285.
11. Xu, Z.; Kang, Y.; Cao, Y. High-Resolution Urban Flows Forecasting With Coarse-Grained Spatiotemporal Data. *IEEE Trans. Artif. Intell.* **2022**, *4*, 315–327.
12. Do, T.H.; Tsiligianni, E.; Qin, X.; Hofman, J.; La Manna, V.P.; Philips, W.; Deligiannis, N. Graph-Deep-Learning-Based Inference of Fine-Grained Air Quality From Mobile IoT Sensors. *IEEE Internet Things J.* **2020**, *7*, 8943–8955. https://doi.org/10.1109/JIOT.2020.2999446.
13. Dun, A.; Yang, Y.; Lei, F. Dynamic graph convolution neural network based on spatial-temporal correlation for air quality prediction. *Ecol. Inform.* **2022**, *70*, 101736. https://doi.org/https://doi.org/10.1016/j.ecoinf.2022.101736.
14. Vandal, T.; Kodra, E.; Ganguly, S.; Michaelis, A.; Nemani, R.; Ganguly, A.R. Deepsd: Generating high resolution climate change projections through single image super-resolution. In Proceedings of the 23rd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 1663–1672.
15. Harilal, N.; Singh, M.; Bhatia, U. Augmented convolutional LSTMs for generation of high-resolution climate change projections. *IEEE Access* **2021**, *9*, 25208–25218.
16. Jin, G.; Sha, H.; Xi, Z.; Huang, J. Urban hotspot forecasting via automated spatio-temporal information fusion. *Appl. Soft Comput.* **2023**, *136*, 110087.
17. Jin, G.; Liu, C.; Xi, Z.; Sha, H.; Liu, Y.; Huang, J. Adaptive Dual-View WaveNet for urban spatial–temporal event prediction. *Inf. Sci.* **2022**, *588*, 315–330.

18. Jin, G.; Sha, H.; Feng, Y.; Cheng, Q.; Huang, J. GSEN: An ensemble deep learning benchmark model for urban hotspots spatiotemporal prediction. *Neurocomputing* **2021**, *455*, 353–367.
19. Jin, G.; Xi, Z.; Sha, H.; Feng, Y.; Huang, J. Deep multi-view graph-based network for citywide ride-hailing demand prediction. *Neurocomputing* **2022**, *510*, 79–94.
20. Wang, Z.; Yue, S.; Song, C. Video-Based Air Quality Measurement With Dual-Channel 3-D Convolutional Network. *IEEE Internet Things J.* **2021**, *8*, 14372–14384.
21. Liu, H.; Ruan, Z.; Zhao, P.; Dong, C.; Shang, F.; Liu, Y.; Timofte, R. Video super-resolution based on deep learning: A comprehensive survey. *Artif. Intell. Rev.* **2022**, *55*, 5981–6035.
22. Jordanova, N.; Jordanova, D.; Tcherkezova, E.; Georgieva, B.; Ishlyamski, D. Advanced mineral magnetic and geochemical investigations of road dusts for assessment of pollution in urban areas near the largest copper smelter in SE Europe. *Sci. Total Environ.* **2021**, *792*, 2613.
23. Qin, X.; Do, T.H.; Hofman, J.; Bonet, E.R.; La Manna, V.P.; Deligiannis, N.; Philips, W. Fine-grained urban air quality mapping from sparse mobile air pollution measurements and dense traffic density. *Remote Sens.* **2022**, *14*, 2613.
24. Qu, K.; Yu, T.; Shi, S.; Chen, Y. Synergetic power-gas flow with space-time diffusion control of air pollutants using a convex multi-objective optimization. *IEEE Trans. Sustain. Energy* **2019**, *11*, 726–735.
25. Zhu, J.; Zhou, X.; Cong, B.; Kikumoto, H. Estimation of the point source parameters by the adjoint equation in the time-varying atmospheric environment with unknown turn-on time. *Build. Environ.* **2023**, *230*, 110029.
26. Manisalidis, I.; Stavropoulou, E.; Stavropoulos, A.; Bezirtzoglou, E. Environmental and Health Impacts of Air Pollution: A Review. *Front. Public Health* **2020**, *8*, 14.
27. Johannessen, C.; Liggio, J.; Zhang, X.; Saini, A.; Harner, T. Composition and transformation chemistry of tire-wear derived organic chemicals and implications for air pollution. *Atmos. Pollut. Res.* **2022**, *13*, 101533.
28. van Rossum, G. Python. 2019. Available online: https://pythoninstitute.org/about-python (accessed on 22 January 2025)
29. Jiang, W.; Gao, W.; Gao, X.; Ma, M.; Zhou, M.; Du, K.; Ma, X. Spatio-temporal heterogeneity of air pollution and its key influencing factors in the Yellow River Economic Belt of China from 2014 to 2019. *J. Environ. Manag.* **2021**, *296*, 113172. https://doi.org/https://doi.org/10.1016/j.jenvman.2021.113172.
30. Yuan, Q.; Qi, B.; Hu, D.; Wang, J.; Zhang, J.; Yang, H.; Zhang, S.; Liu, L.; Xu, L.; Li, W. Spatiotemporal variations and reduction of air pollutants during the COVID-19 pandemic in a megacity of Yangtze River Delta in China. *Sci. Total Environ.* **2021**, *751*, 141820. https://doi.org/https://doi.org/10.1016/j.scitotenv.2020.141820.
31. Deng, C.; Tian, S.; Li, Z.; Li, K. Spatiotemporal characteristics of $PM_{2.5}$ and ozone concentrations in Chinese urban clusters. *Chemosphere* **2022**, *295*, 133813. https://doi.org/https://doi.org/10.1016/j.chemosphere.2022.133813.
32. He, W.; Meng, H.; Han, J.; Zhou, G.; Zheng, H.; Zhang, S. Spatiotemporal $PM_{2.5}$ estimations in China from 2015 to 2020 using an improved gradient boosting decision tree. *Chemosphere* **2022**, *296*, 134003. https://doi.org/https://doi.org/10.1016/j.chemosphere.2022.134003.
33. Chen, G.; Wang, Y.; Li, S.; Cao, W.; Ren, H.; Knibbs, L.D.; Abramson, M.J.; Guo, Y. Spatiotemporal patterns of PM10 concentrations over China during 2005–2016: A satellite-based estimation using the random forests approach. *Environ. Pollut.* **2018**, *242*, 605–613. https://doi.org/https://doi.org/10.1016/j.envpol.2018.07.012.
34. Wu, M.; Huang, J.; Liu, N.; Ma, R.; Wang, Y.; Zhang, L. A Hybrid Air Pollution Reconstruction by Adaptive Interpolation Method. In Proceedings of the SenSys '18: The 16th ACM Conference on Embedded Networked Sensor Systems, Shenzhen, China, 4–7 November 2018, SenSys '18; New York, NY, USA, 2018; pp. 408–409. https://doi.org/10.1145/3274783.3275207.
35. Blanchard, C.; Tanenbaum, S.; Hidy, G. Spatial and temporal variability of air pollution in Birmingham, Alabama. *Atmos. Environ.* **2014**, *89*, 382–391. https://doi.org/https://doi.org/10.1016/j.atmosenv.2014.01.006.
36. Wei, J.; Li, Z.; Lyapustin, A.; Sun, L.; Peng, Y.; Xue, W.; Su, T.; Cribb, M. Reconstructing 1-km-resolution high-quality $PM_{2.5}$ data records from 2000 to 2018 in China: Spatiotemporal variations and policy implications. *Remote Sens. Environ.* **2021**, *252*, 112136. https://doi.org/https://doi.org/10.1016/j.rse.2020.112136.
37. Wei, J.; Li, Z.; Cribb, M.; Huang, W.; Xue, W.; Sun, L.; Guo, J.; Peng, Y.; Li, J.; Lyapustin, A.; et al. Improved 1 km resolution $PM_{2.5}$ estimates across China using enhanced space–time extremely randomized trees. *Atmos. Chem. Phys.* **2020**, *20*, 3273–3289. https://doi.org/10.5194/acp-20-3273-2020.
38. Babaan, J.; Hsu, F.T.; Wong, P.Y.; Chen, P.C.; Guo, Y.L.; Lung, S.C.C.; Chen, Y.C.; Wu, C.D. A Geo-AI-based ensemble mixed spatial prediction model with fine spatial-temporal resolution for estimating daytime/nighttime/daily average ozone concentrations variations in Taiwan. *J. Hazard. Mater.* **2023**, *446*, 130749. https://doi.org/https://doi.org/10.1016/j.jhazmat.2023.130749.
39. Yang, X.; Zhang, Z. An attention-based domain spatial-temporal meta-learning (ADST-ML) approach for $PM_{2.5}$ concentration dynamics prediction. *Urban Clim.* **2023**, *47*, 101363. https://doi.org/https://doi.org/10.1016/j.uclim.2022.101363.
40. Liu, X.; Zhao, J.; Lin, S.; Li, J.; Wang, S.; Zhang, Y.; Gao, Y.; Chai, J. Fine-Grained Individual Air Quality Index (IAQI) Prediction Based on Spatial-Temporal Causal Convolution Network: A Case Study of Shanghai. *Atmosphere* **2022**, *13*, 959. https://doi.org/10.3390/atmos13060959.

41. Ma, P.; Tao, F.; Gao, L.; Leng, S.; Yang, K.; Zhou, T. Retrieval of Fine-Grained $PM_{2.5}$ Spatiotemporal Resolution Based on Multiple Machine Learning Models. *Remote Sens.* **2022**, *14*, 599. https://doi.org/10.3390/rs14030599.
42. Xu, Z.; Kang, Y.; Cao, Y.; Li, Z. Spatiotemporal graph convolution multifusion network for urban vehicle emission prediction. *IEEE Trans. Neural Networks Learn. Syst.* **2020**, *32*, 3342–3354.
43. Zhang, Q.; Han, Y.; Li, V.O.; Lam, J.C. Deep-AIR: A hybrid CNN-LSTM framework for fine-grained air pollution estimation and forecast in metropolitan cities. *IEEE Access* **2022**, *10*, 55818–55841.
44. Hu, K.; Guo, X.; Gong, X.; Hu, Y.; Lin, J.C.W. Spatial-Temporal Air Quality Inference based on Matrix Factorization. In Proceedings of the ICC 2022—IEEE International Conference on Communications, Seoul, Republic of Korea, 16–20 May 2022; pp. 5092–5097. https://doi.org/10.1109/ICC45855.2022.9839163.
45. Hofman Jelleand Do, T.H.; Qin, X.; Rodrigo, E.; Nikolaou, M.E.; Philips, W.; Deligiannis, N.; Manna, V.P.L. Spatiotemporal Air Quality Inference of Low-Cost Sensor Data; Application on a Cycling Monitoring Network. In Proceedings of the Pattern Recognition. ICPR International Workshops and Challenges, Virtual Event, 10–15 January 2021; Del Bimbo, A., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J., Vezzani, R., Eds.; Cham, Switzerland, 2021; pp. 139–147.
46. Marjovi, A.; Arfire, A.; Martinoli, A. High Resolution Air Pollution Maps in Urban Environments Using Mobile Sensor Networks. In Proceedings of the 2015 International Conference on Distributed Computing in Sensor Systems, Fortaleza, Brazil, 10–12 June 2015; pp. 11–20. https://doi.org/10.1109/DCOSS.2015.32.
47. Standard Map. 2024. Available online: https://cmpmap.com/ (accessed on 22 January 2025).
48. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
49. Chan, K.C.; Wang, X.; Yu, K.; Dong, C.; Loy, C.C. BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 4947–4956.
50. Ranjan, A.; Black, M.J. Optical Flow Estimation Using a Spatial Pyramid Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017
51. Liang, Y.; Ouyang, K.; Jing, L.; Ruan, S.; Liu, Y.; Zhang, J.; Rosenblum, D.S.; Zheng, Y. Urbanfm: Inferring fine-grained urban flows. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 3132–3142.
52. Wang, X.; Chan, K.C.; Yu, K.; Dong, C.; Change Loy, C. EDVR: Video Restoration With Enhanced Deformable Convolutional Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, USA, 15–20 June 2019.
53. Chan, K.C.; Zhou, S.; Xu, X.; Loy, C.C. BasicVSR++: Improving Video Super-Resolution With Enhanced Propagation and Alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5972–5981.
54. Yu, J.; Liu, J.; Bo, L.; Mei, T. Memory-Augmented Non-Local Attention for Video Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 17834–17843.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.