TFAN: Temporal-Feature correlations Attention-based Network for Urban Air Quality Prediction using Data Fusion technology

Siyuan MA masiyuan007@qq.com
Henan Academy of Big Data, Zhengzhou University, Zhengzhou, China
School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China

Fan ZHANG zhangfan@ncwu.edu.cn
North China University of Water Resources and Electric Power, Zhengzhou, China

Wanli HOU wlhou123@163.com
Henan Academy of Big Data, Zhengzhou University, Zhengzhou, China
School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China

Yarui LI 2021107247@zut.edu.cn
School of Computer, Zhongyuan University of Technology, Zhengzhou, China

Wei SONG* iewsong@zzu.edu.cn
Henan Academy of Big Data, Zhengzhou University, Zhengzhou, China

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

Air pollution raises a detrimental impact on human health and natural environment. Accurate prediction of air quality is crucial for effective pollution control and mitigation strategies. Numerous existing methods for analyzing the variation tendency of a specific air component primarily focus on its temporal and spatial information, neglecting the potential interactions between different attributes within the same time interval. In this paper, we propose a Temporal-Feature correlations Attention-based deep learning Network (TFAN), which incorporates data fusion technology. TFAN focuses on capturing temporal dependencies, feature correlations, and the potential relationship between temporal-feature through the Attention mechanism, and the data fusion method allows for a comprehensive consideration of multiple factors on prediction. Experimental results conducted using real-world data from Beijing City demonstrate that TFAN outperforms various baseline models in prediction accuracy for multiple pollutants by 10+%.

Keywords: Air Quality; Time Series Prediction; Attention Mechanism; Data Fusion.

1. Introduction

The rapid development of society and cities has not only improved people's quality of life, but also caused a series of environmental issues, especially the air pollution problem. PM2.5 was the fifth-ranking mortality risk factor in 2015 (Cohen et al., 2017). Additionally, prolonged exposure to particulate matter and NO_2 can result in irreversible respiratory diseases (Shi et al., 2016) and significantly increase the risk of death from cardiovascular conditions, thereby reducing life expectancy (Brook et al., 2010). Air quality prediction plays a vital

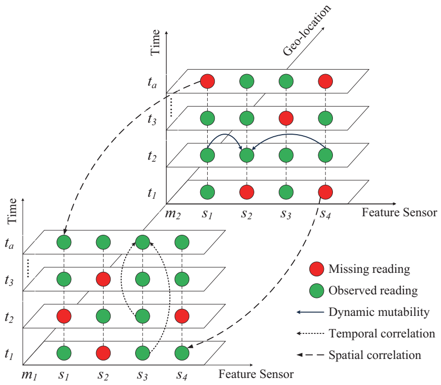


Figure 1: Geo-sensory multivariate time series and the illustration of dynamic interaction

role in safeguarding urban environments and protecting human health. However, forecasting air quality data with Geo-sensory information is challenging due to the following reasons:

1. The dynamic mutability of air quality data. Due to the interaction among air components, changes in a single variable can cause indirect reactions and continuous effects to the remaining variables (Huang et al., 2021) (represented by the solid arrow in Figure 1.). Correspondingly, Qi et al. (2018) deem CO, PM10 and PM2.5 itself the most relevant features for predicting PM2.5.

2. Sensor-level intra-characteristic and external factors. Readings from a specific sensor usually follow a periodicity and changes over time and varies geographically (Zhang et al., 2017). Additionally, sensors' readings are also affected by the external factors such as meteorological data, and the time period of the day (Liang et al., 2018).

3. Complex temporal correlation and unstable spatial correlation. Air quality exhibits a dependency on historical data in different periods (e.g. the short dotted line in Figure 1.), however, due to the significant fluctuation, the role of short-term historical data is greatly weakened; Additionally, Geo-sensory time series vary by locations non-linearly (Yi et al., 2016). Diverse factors make the impact of spatial data uncertain, e.g. the geographical environment and the distance between stations, etc. This uncertainty presents a contradiction to the First Law of Geography (Tobler, 1970). Moreover, due to the maintenance and damage to equipments, the data contains numerous missing values (e.g. the red circle in Figure 1.), which renders the available spatial data unstable. Although various imputing methods exist to fill the gaps (Yi et al., 2016; Guo et al., 2019; Ngueilbaye et al., 2021), biased estimation can lead to error accumulation.

To address the aforementioned challenges, we propose a Temporal-Feature correlations Attention-based Deep Learning Network (TFAN) for urban air quality prediction. The contribution of this paper are outlined as follows:

1. Multi-view attention-based model. We build a multi-view network based on attention mechanism to further improve the prediction accuracy. Attention fully investigates the

correlations between features, timestamps and temporal-feature of each points within multi-variable air quality time series. And the integration of multi-views enables the network to achieve more stable and reliable prediction patterns.

2. Data fusion of external and internal factors. TFAN integrates meteorological data to enhance the influence of external factors, and temporal data to enhance the internal time information. Additionally, Padding data is added as the part to be predicted, enabling each future time step to access multiple sources of information.

3. Result of varying forecast lengths and the dependency of look back window size. We conduct extensive experiments on the datasets constructed with diverse forecast lengths to verify the effectiveness and robustness of TFAN. The sensitivity on look back window size reflects the pollutants' temporal dependence on historical data, and the retention duration in the atmosphere.

Unlike many existing models that focus solely on predicting PM2.5 concentrations, we apply real-world data from Beijing to conduct comprehensive experiments on six pollutants with varies forecast length, aims to provide more convincing verification and evaluation.

2. Related Work

Nowadays, the rising data value prompts people to re-examine massive air quality data from a new perspective, correspondingly, various models are constantly proposed. They can be roughly divided into statistical model, machine learning model, and deep neural network.

Statistical models use numerous historical data and methods such as Clustering and Multiple Regression to analyze potential air quality patterns, e.g. Linear Regression (Dun et al., 2020) and AutoRegressive Integrated Moving Average (ARIMA) (Box and Pierce, 1970). They have extensive capabilities and scalability, and require less training time. However, the methods heavily rely on time series statistical techniques (Huang et al., 2021), and fine-gained raw data is hard to obtain. Furthermore, the above linear methods are inconsistent with the complex nonlinear relations exhibited by air quality attributes.

With the higher accuracy demand in scientific fields, machine learning and deep neural network are receiving increasing attention. And the rapid development of infrastructure sees a concomitant increase in available data. Traditional machine learning models e.g. Support Vector Machines have been used to predict air quality index and concentration (Castelli et al., 2020). Liu et al. (2019) constructs a hybrid model based on Back Propagation (BP) and optimization algorithms to predict air quality. Multi-Layer Perceptron effectively models air quality data by adding nonlinear activation. Machine learning methods achieve good results in forecasting air pollutants, but they are hard to deal with the high-dimensional nonlinear long time series problem in the prediction process (Ma et al., 2020).

On this basis, researchers propose deep learning methods to address nonlinear long time series problems. The models based on LSTM (Hochreiter and Schmidhuber, 1997) is improving gradually in predicting air quality (Al-Janabi et al., 2020), which use long and short term memory to learn long-term dependence of time series. DAL (Qi et al., 2018) embeds feature selection and semi-supervised learning in different layers, and uses unlabeled spatio-temporal information and data fusion to improve the interpolation and prediction accuracy. GeoMAN (Liang et al., 2018) predicts a geo-sensor readings over several future hours by using a multi-level attention-based RNN, which considers multiple sensors' data, meteorological data, and spatial data. Bhattacharyya et al. (2021) propose a transformer-based model called cosSquareformer and a non-linear re-weighting attention mechanism, which is firstly applied to multi-variate pollutant forecasting tasks. Zheng et al. (2015) use four modules to model local factors, global factors, spatio-temporal meteorological data and inflection point respectively. However, most existing spatial modeling solutions neglect the inherent multi-scale spatial correlations, and they may encounter over-fitting problem because of limited available data compared to the complicated model size (Wu et al., 2019).

Transformer is deemed to be a very successful sequence modeling architecture, and demonstrates unmatched performance in a variety of applications. WU et al. propose a deep learning model named Autoformer (Wu et al., 2021), which is empowered with progressive decomposition capacities for complex time series, and they design Auto-Correlation mechanism based on the series periodicity to conducts the dependencies discovery and representation. Informer (Zhou et al., 2021) builds the Prob Sparse self-attention mechanism to reduce the computational complexity, and uses generative decoder to predict the time series through a one-step forward operation, which greatly improves the reasoning speed.

3. Methodology

3.1. Data Fusion

Data fusion system is able to reemerge the full view of an observed phenomenon by transforming data into a modality with more value and higher quality (Meng et al., 2020). It can also provide specific benefits for some application contexts (Khaleghi et al., 2013), correspondingly, an effective fusion can reveal the positive effects of potential features, thereby improving the pollutants forecast accuracy. As shown in Figure 2., *Sequence Length* represents the look back window size and *Predicted Length* stands for the forecast length.

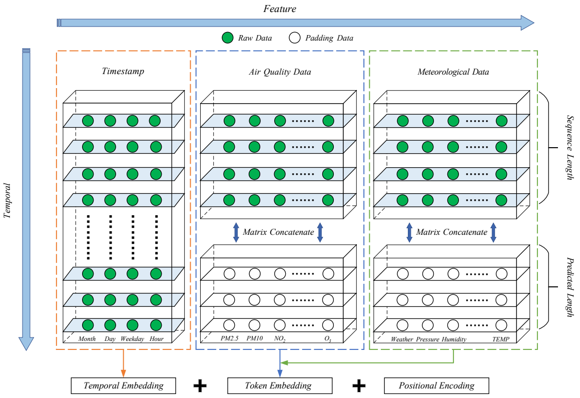


Figure 2: Data fusion of the network. The X-axis stands for multiple attributes, and the Y-axis represents the time dimension.

1. Meteorological data. Meteorological conditions significantly affect air quality, e.g. rainy weather, in particular, is widely recognized for its ability to freshen the air, which inspires us to obtain corresponding region meteorological data to consider its impact on prediction. TFAN merges it with the air quality data for *Token Embedding*.

2. Timestamp data. Self-attention is permutation invariant and "anti-order" to some extent (Zeng et al., 2022). TFAN maps discrete timestamp data which contains month, day, weekday and hour into a high-dimensional vector through embedding operation to enhance the representation of time information.

3. Padding. Vanilla Transformer's decoder splices token and padding as decoder input, where the token marks the starting position of the sequence, while the padding represents the unknown portion. Informer and Autoformer also uses a concatenated vector of token and padding as decoder input, the distinction from Vanilla is only that the token acts as prior knowledge, which is derived from a subsequence of the encoder input. It can be found that the padding often participates in correlation calculation as the data to be predicted. Hence, TFAN directly concatenates the input matrix with padding data. This allows Attention to calculate the correlation between the unknown points and all timestamps and all features.

4. Positional encoding. TFAN uses positional encoding (Vaswani et al., 2017) applied in Vanilla Transformer to further add position information.

3.2. Attention Mechanism

The self-attention mechanism used in Transformer-based models typically operates on the same input for $Query$, Key, $Value$ matrix, which motivates the calculations of potential associations within the same dimension. In addition, most time series have obvious periodicity and time dependence, and their previous data holds significant reference value for prediction. However, the air quality data tends to have strong volatility, which extremely diminishes the effect of its neighboring value. Based on the above cognition, TFAN utilizes Self-Attention to concentrate on timestamp dimension and feature dimension separately while considering the prior data points. Then carries out the co-attention calculation for $Temporal$-$Feature$ and $Feature$-$Temporal$ after Self-Attention, where $Query$, Key and $Value$ are from different *View*'s self-attention output. As depicted in Figure 3, the dotted line box represents the quantified correlation degree between the first time step and the first feature. Similarly, the short dotted line box represents the same between the first unknown time step and the first feature. This constructs a constraint between time and features, enables the network to potentially handle the volatility and non-stationarity observed in air quality data more effectively, and not restricted to the limited change forecast.

3.3. Model Architecture

As shown in Figure 4, Input data is divided into timestamp data *x_stamp*, and numerical data *x_data*. x_data includes air quality data, meteorological data, and Padding. There is significant numerical distribution gap between attributes (e.g. the mean value of CO is 1.2 and that of PM10 is 126.2), TFAN incorporates Batch Normalization to reduce the

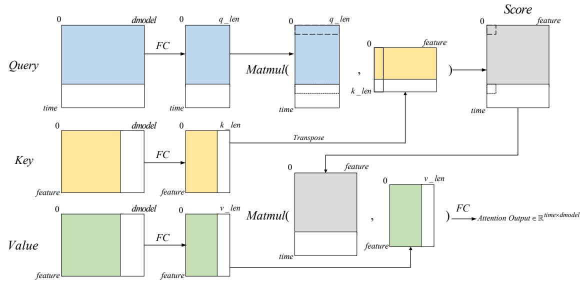


Figure 3: Details of Temporal-Feature Attention mechanism. The colorful part stands for value data, and the blank part of the rectangle is Padding.

impact of the gap and decrease the sensitivity of the initial parameters. Each value of *x_stamp* has a specific meaning, which will be mapped to a high-dimensional vector using a word embedding optional. The *Embedding Layer* fuses multiple information (*x_stamp* is not applied to *Feature View*), as shown in Equation 1 and 2, and *Token_embedding* performs high-dimensional representation of low-dimension vectors. *Dropout* is embedded between layers to effectively reduce the overfitting degree of the network (Srivastava et al., 2014).

$$Embedding = Temporal\ Embedding + Token\ Embedding + Positional\ Encoding \tag{1}$$

$$Token\ Embedding = Dropout(Conv1d(x\_data)) \tag{2}$$

Multi-Head Attention (*MHA*) in both *Views* focus on the correlations between dimensions separately. Since *MHA* is all linear mapping, the *Feedforward Layer* after it adds nonlinear transformation, which enhances the ability and accelerates the speed of fitting. As shown in Equation 3 and 4, Layer Normalization (*LN*) operation is used to normalize the layer parameter distribution. *Residual Representations* prevents gradient disappearance caused by excessive depth of the model (He et al., 2016), and ReLU adds nonlinear activation (Nair and Hinton, 2010) to enhance fitting ability, Encoder cycles N times.

$$MHA\ Layer = LN(Residual(Dropout(MHA(x)))) \tag{3}$$

$$Feed\ Forward = LN(Residual(Conv1d(Dropout(ReLU(Conv1d(Dropout(x))))))) \tag{4}$$

So far, the hidden output of the two views are denoted as *x_time* and *x_feature* respectively. Then TFAN works on *Temporal-Feature correlation(TFC)* and *Feature-Temporal correlation(FTC)* through *MHA*. Their specific processes are shown in Equation 5 and 6.

$$TFC = MHA(Q = x\_time,\ K = x\_feature,\ V = x\_feature) \tag{5}$$

$$FTC = MHA(Q = x\_feature,\ K = x\_time,\ V = x\_time) \tag{6}$$

Finally, the output of the two *Views* is mapped to the output dimension through a fully connected layer after dimension reduction, and are still represented as *x_time* and *x_feature*. *MEAN* in Figure 4. represents the mean value of target attribute in the look back window interval. The final prediction result is obtained through the weighted summation layer, as shown in Equation 7, where $\{w_0,\ w_1,\ w_2,\ bias\}$ is the weight and bias.

$$Output = w_0 \times x\_time + w_1 \times x\_feature + w_2 \times mean + bias \tag{7}$$

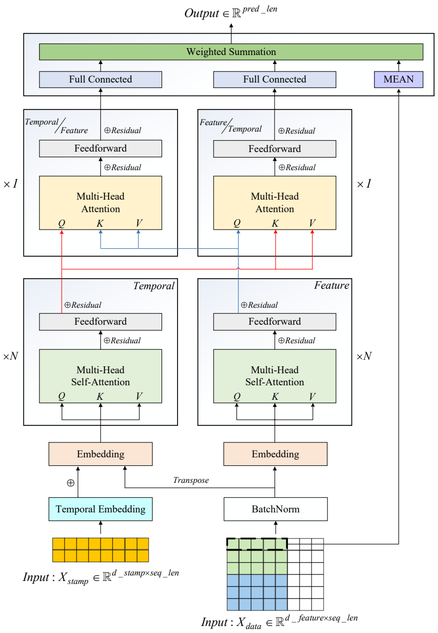


Figure 4: Architecture of TFAN

4. Experiment

4.1. Dataset

As a universal norm, the test dataset is taken from a certain time period or season of one geographical position. This truncates the seasonality of the data, and leads to a gap between the model effect and the reality. We consider the station as the unit and randomly select the annual data of several stations as the training set, verification set or test set, and the sets are disjoint. Zero-mean normalization is also applied to all data.

Table 1: Distributions of different pollutants in the data

Pollutions	Mean	25%	50%	75%
PM2.5	83.5	23.0	60.0	118.0
PM10	126.2	50.2	101.0	167.5
NO_2	50.6	21.0	43.1	73.2
CO	1.2	0.5	0.8	1.5
O_3	58.0	9.4	45.1	84.3
SO_2	16.9	3.0	8.7	21.3

Table 2: IAQI and the corresponding pollutants item concentration limits

IAQI	Level	PM2.5	NO_2	CO	SO_2
0	Excellent	0	0	0	0
50	Good	35	40	2	150
100	Light	75	80	4	500
150	Moderate	115	180	14	650
200	Heavy	150	280	24	800
300+	Serious	250+	565+	36+	-

The experiment uses authentic public data collected from 37 air quality stations and 17 meteorological stations in Beijing. The data includes a total of 12 attributes e.g. PM2.5, PM10, weather etc. (Zheng et al., 2015, 2014, 2013). The missing values in the raw data are filled with adjacent values or zero. Table 1 shows the numerical distribution of various pollutants' concentrations, and Table 2 illustrates the corresponding relationship between pollutant concentrations and the Individual Air Quality Index (IAQI) defined in literature (MEP, 2012). See the footer[1] for the details and code.

4.2. Experiment and Parameters Settings

The experiment implements the early stop strategy to prevent the overfitting. We run 5 times with different random seeds and report the MAE and RMSE metrics to evaluate the prediction effect (Botchkarev, 2018). We use L1 loss to calculate the gap between output and the ground truth. The parameters settings for the experiment are shown in Table 3.

Table 3: Parameters settings

Parameter	Value	Parameter	Value	Parameter	Value
Batch size	32	h	64	LR/LR of weighted summation	0.0001/0.001
d_model	512	N	2	$\{w1,\ w2,\ w3,\ bias\}$	$\{0.33,\ 0.33,\ 0.33,\ 0\}$
d_hidden	512	dropout	0.05	Patience of early stop	10
q,k,v	8	Optimizer	Adam	Loss function	L1Loss

4.3. Result and Analysis

4.3.1. Prediction of pollutions concentration

We select the statistics-based model (*Closest Repeat*), concise machine learning model (*LR*, *MLP*, *KNN* with k=7, *RF* with 100 trees), outstanding time series prediction deep learning model (*LSTM*, *Autoformer* (Wu et al., 2021), *Informer* (Zhou et al., 2021)), and neural network (cosSquareformer (Bhattacharyya et al., 2021)) specially for air quality prediction as benchmarks. The experiment predict the pollutants' concentration for the next {24, 48, 72, 96, 120, 144} hours. Because of the discontinuity in time, different predicted lengths will cause large differences in the number of samples and the distribution among data sets. Therefore, the errors of models with different predicted length not only reflect the prediction effectiveness, but also indicate their robustness. Table 4 shows the results for the forecast length in {1, 3, 5} days. The optimal result is highlight in bold, and the suboptimal is underlined. IMP. represents the improvement achieved by surpassing suboptimal results. Please refer to Appendix A for complementary results. The following points can be deduced:

[1] The code has been published on Github

Table 4: Comparison of pollutants concentration prediction precision of various models

Seq_len=168h		IMP.	OURS		cosSquareformer		Informer		MLP		LSTM	
Pollutant	Pred_len	MAE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
PM2.5	24	10.87%	**12.203**	**19.696**	15.832	23.426	13.692	21.608	16.850	25.108	21.699	31.918
	72	18.29%	**11.174**	**16.819**	14.449	21.139	13.676	20.193	18.540	26.190	21.180	29.293
	120	12.97%	**11.344**	**17.074**	13.988	20.522	13.034	19.148	19.919	31.133	23.632	32.083
PM10	24	7.81%	**32.846**	**50.654**	37.997	59.135	35.629	56.260	37.938	57.956	39.168	65.808
	72	12.40%	**30.973**	**45.919**	40.768	59.565	35.356	52.605	42.463	60.130	42.316	67.799
	120	11.93%	**30.634**	**42.407**	35.018	49.366	34.784	47.818	42.153	55.935	43.122	61.581
NO_2	24	12.89%	**11.302**	**16.010**	13.395	18.439	12.974	17.768	15.070	20.575	13.663	19.206
	72	10.77%	**11.793**	**16.293**	13.302	18.345	13.216	17.805	15.784	21.273	15.179	20.919
	120	11.47%	**11.767**	**16.330**	13.291	18.167	13.436	18.200	16.271	21.833	14.308	19.143
CO	24	15.89%	**0.180**	**0.310**	0.225	0.441	0.214	0.361	0.245	0.385	0.264	0.423
	72	16.26%	**0.170**	**0.263**	0.210	0.403	0.203	0.324	0.265	0.384	0.271	0.388
	120	11.52%	**0.169**	**0.254**	0.191	0.328	0.196	0.293	0.273	0.384	0.269	0.378
O_3	24	6.56%	**15.893**	**21.553**	17.009	23.668	18.401	25.604	20.915	28.400	23.046	32.386
	72	8.05%	**16.415**	**22.208**	17.852	24.522	20.100	27.624	23.439	31.425	25.799	35.071
	120	5.00%	**16.881**	**22.730**	17.769	24.206	20.355	27.219	24.853	33.610	26.287	34.694
SO_2	24	10.47%	**3.454**	**5.726**	4.075	7.095	3.858	6.522	4.425	6.996	5.124	8.331
	72	7.15%	**3.441**	**5.890**	4.187	6.859	3.706	6.330	4.580	6.993	5.290	8.820
	120	5.04%	**3.148**	**5.189**	3.715	7.166	3.315	5.568	4.705	7.190	5.421	9.609

1. TFAN achieves optimal prediction accuracy for various air pollutants and significantly surpasses the second-place. It also maintains the prediction accuracy at various predicted lengths. Combined with Table 2, the MAE is relatively small compared with the differentiation of IAQI levels, which demonstrates the practical significance and feasibility of TFAN for pollution prevention and early warning. *Informer* also shows excellent performance by achieving ideal prediction results on SO_2 and NO_2. In addition, *cosSquareformer* achieves a similar effect with TFAN on O_3.

2. Compared with concise machine learning models and statistical models, sophisticated deep learning models perform better. *MLP* has the best results among the machine learning baselines, slightly surpassing *Autoformer*, but still inferior to Transformers.

3. Compared with the linear method, the nonlinear method performs better, e.g. the comparison between *MLP* and *LR*, which is partially attributed to the complexity and instability of air quality. The models considering multi-variable factors, i.e. multivariate models, achieve better and more stable effect than the models considering univariate factor, i.e. univariate models. The former could acquire more available information to further learn the target's change pattern, while the latter only considers temporal changes and overlooks the potential interactions between pollutants.

Figure 5. shows the MAE result for six pollutants using various models, inference:

1. *Repeat* has the largest error among all baselines. Its result curve is parallel to the X-axis, to some extent, the Euclidean Distance between it and the real curve reflects the

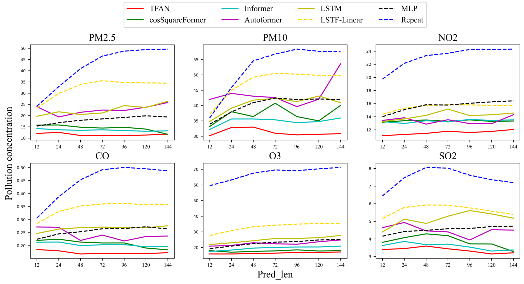


Figure 5: MAE result for six pollutions using various models at varying forecast length. The x-axis represents the predicted length in {0.5d, 1d, 2d, 3d, 4d, 5d and 6d}, and the Y-axis represents the MAE result. The dotted lines represent univariate models, and the solid line represents the multi-variable models.

jitter degree or prediction difficulty. Despite the potential variations due to sample differences, the pollutants (except SO_2) errors see a concomitant increment with longer predicted length, i.e. prediction difficulty is positively correlated with output length.

2. Univariate models exhibit a consistent upward trend. They could learn more universal patterns and overall trends, but ignore the details of temporal changes. Multivariate models such as TFAN, *Informer* and *cosSquareformer* show a stable performance across diverse dataset sizes, while *Autoformer* and *LSTM* show a significant gap.

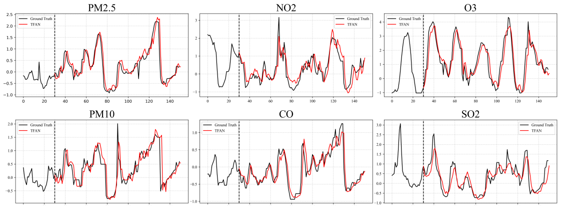


Figure 6: General fit of TFAN on six pollutants which from the same sample in test dataset. The X-axis represents the time point and the Y-axis represents the value after zero-mean normalization. The vertical dotted line indicates the starting position.

Figure 6. shows the general fit of TFAN in future 120h. It can be concluded that:

1. TFAN demonstrates a strong predictive capability across various attributes and effectively handles different patterns of curve changes, e.g. wave crest, wave trough, periodicity, suddenly drop and sharp protrusion. All pollutants' curves exhibit rapid and steep changes without clear periodicity except for O_3.

2. The similarity between various pollutants' curves, such as PM2.5, PM10, CO and NO_2, is relatively high. This intuitively reflects the potential dependencies between air components and the necessity of data fusion in the air quality prediction task.

4.3.2. Ablation

As shown in Table 5, No_Padding represents TFAN without fusing Padding, No_TF replace *Temporal-Feature* Attention with vanilla self-attention, and No_Mean drops the MEAN *View*. NUM counts the optimal number of datasets obtained by each model. ARM represents the average ranking of prediction errors, the lower, the better, and it can be calculated by $ARM = \frac{1}{N}\sum_{i}^{N} Rank_i$, where N stands for the dataset number and $Rank_i$ represents the ranking on the i-th dataset. NUM indicates TFAN's excellent generality, and its ability to get higher accuracy and lower variance. ARM indicates the TFAN's stability and robustness on predictions, it always achieves the best or close rank to the best.

Table 5: Ablation of TFAN

Seq_len=168		TFAN		No_Padding		No_TF		No_Mean	
Pollutant	Pred_len	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
PM2.5	24	12.203	19.696	12.077	**19.134**	**12.011**	19.460	12.236	19.864
	72	11.174	16.819	11.314	16.918	**11.111**	**16.789**	11.372	17.115
	120	11.344	**17.074**	11.402	17.091	11.534	17.307	**11.331**	17.202
PM10	24	**32.846**	**50.654**	33.086	51.252	33.657	52.427	33.529	53.223
	72	**30.973**	**45.919**	31.948	46.869	31.642	46.804	32.826	48.449
	120	**30.634**	**42.407**	31.740	43.890	31.259	43.053	31.649	44.280
NO_2	24	**11.302**	**16.010**	11.436	16.182	11.364	16.106	12.271	16.706
	72	11.793	16.293	11.782	16.259	**11.612**	**16.161**	12.553	16.899
	120	**11.767**	**16.330**	12.024	16.571	11.989	16.497	12.513	16.781
CO	24	**0.180**	**0.310**	0.181	0.313	0.181	0.315	0.189	0.320
	72	0.170	**0.263**	0.171	0.264	**0.169**	0.264	0.181	0.272
	120	**0.169**	**0.254**	0.172	0.256	0.170	**0.254**	0.179	0.261
O_3	24	**15.893**	**21.553**	15.989	21.751	15.932	21.659	15.938	21.655
	72	**16.415**	**22.208**	16.639	22.513	16.529	22.297	16.528	22.340
	120	**16.881**	**22.730**	16.979	22.850	16.983	22.835	16.947	22.805
SO_2	24	3.454	**5.726**	3.504	5.884	3.445	5.901	**3.398**	5.804
	72	3.441	5.890	3.392	5.857	3.419	**5.710**	**3.340**	**5.710**
	120	3.148	**5.189**	3.144	5.310	**3.143**	5.229	3.155	5.239
NUM		**10**	**14**	0	1	5	4	3	1
ARM		**1.778**	**1.444**	2.889	2.833	2.167	2.222	3.111	3.333

4.3.3. Sensitivity analysis of sequence length

As shown in Figure 7., $rRMSE{=}RMSE/RMSE_{baseline}$, where the baseline derives from 120h. When Pred_len=24h, the error of PM10 and SO_2 is positively correlated with the sequence length, while PM2.5, NO_2 and CO have an obvious downward trend. When

Pred_len=120h, all pollutants' curves have a noticeable downward trend, while NO_2, O_3 and PM2.5 decline steadily. Overall, PM2.5 and NO_2 tend to perform better with longer sequence lengths, which aligns with the findings reported by Bhattacharyya et al. (2021).

4.3.4. Comparison with LSTF-Linears

Transformers show remarkable capability in sequential modeling task. However, Zeng et al. (2022) think that the Attention mechanism, which serves as the core component of Transformers, will inevitably lose time information. They propose a simple model called *LSTF-Linear*, composed of a single fully connected layer (with variants named *LSTF-NLinear* and *LSTF-DLinear*), outperforms Transformers on many benchmarks, and raises queries about Transformer-based models in long time series forecasting task. Nevertheless, the result in Table 4 shows that TFAN still maintains high accuracy, and is far higher than *LSTFs*.

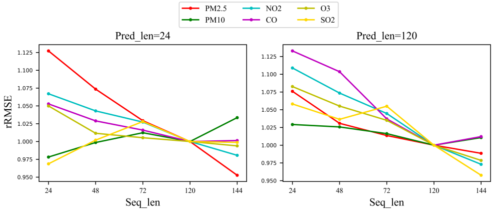
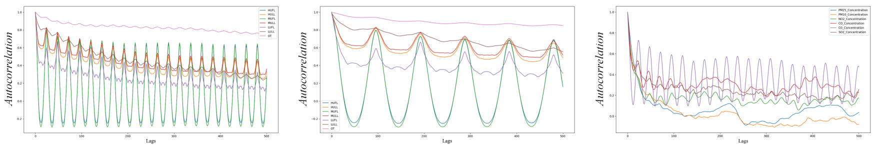


Figure 7: Relative RMSE for different sequence lengths with respect to the 120h baseline for all air pollutions with TFAN. The X-axis stands for the predicted length.

ZENG et al. conduct extensive single-dimensional time series prediction experiments on ETT (Electric Transformer Temperature) dataset (Zhou et al., 2021) and *LSTF-Linear* achieves SOTA. Therefore, we Visualized the autocorrelation of attributes in ETT to explore their temporal dependency and periodicity. As shown in Figure 8. Observe a) and b), most features' autocorrelations are high and decreases slowly with the increase of lags, and the curves have obvious, long-term and stable periodicity, which makes the simple linear-based network competent to the task. On the contrary, the redundant parameters of Transformers can result in overfitting and loss of time information. It is observed from c) that, the autocorrelation of all attributes (except O_3) are low and decreases sharply in the 0 to 50 lags. Moreover, there is no significant long or short term periodicity, PM2.5 and PM10 even reject the assumption of autocorrelation within the 95% confidence interval. It is obviously harder for a simple linear layer to learn the data distribution. However, sophisticated architecture, nonlinear modeling and numerous learnable parameters empower Transformer to explore the correlation between time and features, and overcome complexity and uncertainty.



a) Autocorrelation of ETTh1 *b) Autocorrelation of ETTm1* *c) Autocorrelation of Air Quality*

Figure 8: Autocorrelation of different datasets. The X-axis indicates the lags, and the Y-axis represents the degree of autocorrelation.

5. Conclusion

In this paper, we propose TFAN, which incorporates Attention mechanism and Data Fusion technology for Urban Air Quality Prediction task. The network deeply examines the internal information and co-correlation within multivariate air quality data through Attention. Data Fusion is applied to comprehensively consider the impact of numerous factors on target's variation mode. The comparison results demonstrate the high accuracy of TFAN in prediction tasks, and the ablation sufficiently demonstrates the motivation for design choices. The general fit experiment intuitively reflects the potential dependencies between air attributes and the necessity of data fusion. The sensitivity analysis preliminarily reflects the pollutants' temporal dependence, and the retention duration in the atmosphere. These analyses validate the effectiveness and robustness of TFAN in capturing and utilizing such dependencies for accurate air quality prediction.

Acknowledgments

The Key Scientific Research Projects of Colleges and Universities in Henan Province under Grant No.22A520010. The International Cooperation Project, Department of Science & Technology, Henan Province, China, under Grant No.172102410065.

References

Samaher Al-Janabi, Mustafa Mohammad, and Ali Al-Sultan. A new method for prediction of air pollution based on intelligent computation. *Soft Computing*, 24(1):661–680, 2020.

Mayukh Bhattacharyya, Sayan Nag, and Udita Ghosh. Deciphering environmental air pollution with large scale city data. *arXiv preprint arXiv:2109.04572*, 2021.

Alexei Botchkarev. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *Statistics*, 2018.

George EP Box and David A Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, 65(332):1509–1526, 1970.

Robert D Brook, Sanjay Rajagopalan, C Arden Pope, Jeffrey R Brook, and Joel D Kaufman. Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the american heart association. *Circulation*, 121(21):2331–2378, 2010.

Mauro Castelli, Fabiana Martins Clemente, Aleš Popovič, Sara Silva, and Leonardo Vanneschi. A machine learning approach to predict air quality in california. *Complexity*, 2020, 2020.

Aaron J Cohen, Michael Brauer, Richard Burnett, H Ross Anderson, Joseph Frostad, Kara Estep, Kalpana Balakrishnan, Bert Brunekreef, Lalit Dandona, Rakhi Dandona, et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015. *The lancet*, 389(10082):1907–1918, 2017.

Meng Dun, Zhicun Xu, Yan Chen, and Lifeng Wu. Short-term air quality prediction based on fractional grey linear regression and support vector machine. *Mathematical problems in engineering*, 2020, 2020.

Zijian Guo, Yiming Wan, and Hao Ye. A data imputation method for multivariate time series based on generative adversarial network. *Neurocomputing*, 360:185–197, 2019.

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

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Wei Huang, Tianrui Li, Jia Liu, Peng Xie, Shengdong Du, and Fei Teng. An overview of air quality analysis by big data techniques: Monitoring, forecasting, and traceability. *Information Fusion*, 75:28–40, 2021.

Bahador Khaleghi, Alaa Khamis, Fakhreddine O. Karray, and Saiedeh N. Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28–44, JAN 2013. ISSN 1566-2535.

Yuxuan Liang, Songyu Ke, Junbo Zhang, Xiuwen Yi, and Yu Zheng. Geoman: Multi-level attention networks for geo-sensory time series prediction. In *IJCAI*, volume 2018, pages 3428–3434, 2018.

Hui Liu, Zhu Duan, and Chao Chen. A hybrid framework for forecasting pm2.5 concentrations using multi-step deterministic and probabilistic strategy. *Air Quality Atmosphere Health*, 12(7), 2019.

Jun Ma, Zheng Li, Jack CP Cheng, Yuexiong Ding, Changqing Lin, and Zherui Xu. Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network. *Science of The Total Environment*, 705:135771, 2020.

Tong Meng, Xuyang Jing, Zheng Yan, and Witold Pedrycz. A survey on machine learning for data fusion. *Information Fusion*, 57:115–129, MAY 2020. ISSN 1566-2535.

MEP. Technical regulation on ambient air quality index (on trail). *HJ 633-2012*, 2012.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

Alladoumbaye Ngueilbaye, Hongzhi Wang, Daouda Ahmat Mahamat, and Sahalu B Junaidu. Modulo 9 model-based learning for missing data imputation. *Applied Soft Computing*, 103:107167, 2021.

Zhongang Qi, Tianchun Wang, Guojie Song, Weisong Hu, Xi Li, and Zhongfei Zhang. Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality. *IEEE Transactions on Knowledge and Data Engineering*, 30(12):2285–2297, 2018.

Liuhua Shi, Antonella Zanobetti, Itai Kloog, Brent A Coull, Petros Koutrakis, Steven J Melly, and Joel D Schwartz. Low-concentration pm2. 5 and mortality: estimating acute and chronic effects in a population-based study. *Environmental health perspectives*, 124 (1):46–52, 2016.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34:22419–22430, 2021.

Zhiyuan Wu, Yue Wang, and Lin Zhang. Msstn: Multi-scale spatial temporal network for air pollution prediction. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1547–1556. IEEE, 2019.

Xiuwen Yi, Yu Zheng, Junbo Zhang, and Tianrui Li. St-mvl: filling missing values in geosensory time series data. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, 2016.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *arXiv preprint arXiv:2205.13504*, 2022.

Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. 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*, pages 1436–1444, 2013.

Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):1–55, 2014.

Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. Forecasting fine-grained air quality based on big data. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2267–2276, 2015.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

Appendix A. Complementary result of Table 4.

Table 6: Comparison of pollutants concentration prediction precision of various models

Seq_len=168		Autoformer		Reapeat		LR		LSTF-NLinear		LSTF_DLinear		KNN		RF	
Pollutant	Pred_len	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
PM2.5	24	19.403	28.380	32.893	50.240	29.844	41.646	30.618	43.765	29.680	41.544	13.065	21.195	17.917	26.929
	72	22.472	30.404	46.527	61.578	35.540	45.814	37.238	48.963	35.385	45.722	12.673	19.889	19.139	26.529
	120	23.739	31.507	49.403	63.617	34.584	44.105	36.633	48.145	34.469	44.072	13.580	22.426	19.974	27.280
PM10	24	43.968	66.446	45.908	87.272	45.008	78.609	44.6961	78.1402	44.5791	78.3874	37.531	55.642	38.734	61.888
	72	42.615	63.896	56.801	91.395	50.511	79.402	50.2767	79.441	49.9515	79.1418	39.798	56.412	44.382	63.628
	120	42.082	55.688	57.712	80.547	49.847	68.527	49.6783	68.5402	49.21	67.9288	41.334	55.584	43.490	56.750
NO2	24	13.849	18.964	22.146	32.514	15.310	21.843	15.354	21.869	15.321	21.858	14.357	19.175	16.354	21.098
	72	13.548	18.573	23.678	34.231	15.827	22.301	15.857	22.331	15.820	22.298	14.803	19.703	16.851	21.317
	120	12.937	17.831	24.267	33.355	15.742	22.156	15.809	22.208	15.753	22.171	14.916	20.058	16.613	21.026
CO	24	0.271	0.416	0.388	0.605	0.331	0.501	0.341	0.518	0.328	0.499	0.221	0.385	0.257	0.417
	72	0.241	0.360	0.492	0.682	0.360	0.523	0.372	0.538	0.360	0.523	0.218	0.349	0.266	0.385
	120	0.235	0.333	0.495	0.659	0.357	0.495	0.371	0.514	0.356	0.494	0.223	0.351	0.271	0.378
O3	24	21.359	29.339	63.231	84.934	30.715	41.747	30.8984	42.0025	30.7041	41.7472	18.943	26.685	25.141	33.287
	72	22.386	30.499	69.579	90.984	34.256	45.976	34.5109	46.2865	34.2779	46.009	19.749	27.265	28.128	36.692
	120	23.684	31.276	70.274	90.635	35.332	46.625	35.5986	47.0856	35.3156	46.6194	20.962	29.351	29.272	38.325
SO2	24	4.936	7.876	7.476	13.118	5.787	10.009	5.8966	10.081	5.7681	9.9842	4.014	6.600	4.859	7.460
	72	4.401	7.569	8.017	13.392	5.918	10.321	6.0271	10.2832	5.905	10.3064	3.835	6.185	4.949	7.384
	120	4.528	7.261	7.367	12.192	5.574	9.635	5.7155	9.5798	5.5701	9.6304	3.738	6.384	4.765	7.348