Decoupled Invariant Attention Network for Multivariate Time-series Forecasting

Haihua Xu[1,2], **Wei Fan**[3], **Kun Yi**[4] and **Pengyang Wang**[1,2†]
[1]Department of Computer and Information Science, University of Macau, China
[2]The State Key Laboratory of Internet of Things for Smart City, University of Macau, China
[3]University of Central Florida, USA
[4]Beijing Institute of Technology, China
{yc37901, pywang}@um.edu.mo, weifan@ucf.edu, yikun@bit.edu.cn

Abstract

To achieve more accurate prediction results in Time Series Forecasting (TSF), it is essential to distinguish between the valuable patterns (invariant patterns) of the spatial-temporal relationship and the patterns that are prone to generate distribution shift (variant patterns), then combine them for forecasting. The existing works, such as transformer-based models and GNN-based models, focus on capturing main forecasting dependencies whether it is stable or not, and they tend to overlook patterns that carry both useful information and distribution shift. In this paper, we propose a model for better forecasting time series: Decoupled Invariant Attention Network (DIAN), which contains two modules to learn spatial and temporal relationships respectively: 1) Spatial Decoupled Invariant-Variant Learning (SDIVL) to decouple the spatial invariant and variant attention scores, and then leverage convolutional networks to effectively integrate them for subsequent layers; 2) Temporal Augmented Invariant-Variant Learning (TAIVL) to decouple temporal invariant and variant patterns and combine them for further forecasting. In this module, we also design Temporal Intervention Mechanism to create multiple intervened samples by reassembling variant patterns across time stamps to eliminate the spurious impacts of variant patterns. In addition, we propose Joint Optimization to minimize the loss function considering all invariant patterns, variant patterns and intervened patterns so that our model can gain a more stable predictive ability. Extensive experiments on five datasets have demonstrated our superior performance with higher efficiency compared with state-of-the-art methods.

1 Introduction

Multivariate time series (MTS) forecasting has been increasingly important in a wide range of real-world applications, such as traffic flow analysis [Yu *et al.*, 2017], weather estimation [Zheng *et al.*, 2015], energy consumption planning [Becali *et al.*, 2008], etc. A primary challenge in MTS forecasting is to capture the core inter-series (spatial) correlations and intra-series (temporal) dependencies simultaneously [Cao *et al.*, 2020a; Bai *et al.*, 2020]. Despite advancements, accurate forecasting remains elusive due to the inherent complexities, shifted fluctuations, and noise in time series signals, which pose challenges for model learning [Fan *et al.*, 2023].

Existing literatures have tried different strategies to extract the *core patterns* of time series for accurate forecasting. Some researchers have focused on periodicity modeling to find generalizable periods beneficial for forecasting [Fan *et al.*, 2022; Jiang *et al.*, 2022]; other works priorly define and extract an overall trend of raw series, and jointly learn the trend and the left parts [Wu *et al.*, 2022; Zeng *et al.*, 2023]; besides, [Woo *et al.*, 2022a; Wang *et al.*, 2022] disentangle time series into seasonal and trend and learn their representations respectively. Existing studies typically rely on manually specified core patterns (e.g., periods, trends) of time series based on certain assumptions. Because of focusing solely on these pre-defined patterns and neglecting dependencies on other signals, the performance of multivariate forecasting is significantly compromised.

Inspired by previous works of time series [Lim and Zohren, 2021] and views of domain generalization [Zhou *et al.*, 2022a], we generally categorize time series into two types of patterns for forecasting: (i) *invariant patterns*, the regular and constant patterns of time series that are easy to illustrate, such as regular seasonal or trend signals; (ii) *variant patterns*, the volatile and changeable patterns of time series that are hard to capture, such as useful information mixed with noise. Instead of directly specifying the invariant patterns for learning as [Fan *et al.*, 2022; Wu *et al.*, 2022; Woo *et al.*, 2022a], we aim to build a data-driven approach to adaptively distinguish the invariant patterns from the variant patterns in a given raw time series. After that, we can conduct spatial-temporal modeling for multivariate time series from a disentangled learning perspective. However, two challenges have arisen in achieving this goal:

- *Challenge I: Adaptive decoupling of invariant and variant patterns presents difficulties.* First, these patterns are often intertwined due to the non-stationary nature of time series data, resulting in ambiguous and ever-changing boundaries. Moreover, multivariate time series (MTS) encompasses both spatial and temporal di-

†Corresponding author.

mensions, each containing their own invariant and variant patterns. Simultaneously considering these patterns across the Spatiotemporal dimensions further complicates the decoupling process.

- *Challenge II: In the temporal dimension, distribution shift are notably pronounced.* This heightened variability introduces a substantial perturbation, making it increasingly challenging to discern and differentiate between invariant and variant patterns over time. Furthermore, these shift can adversely impact subsequent forecasting tasks, leading to performance drift.

To address the above challenges, in this paper, we propose **D**ecoupled **I**nvariant **A**ttention **N**etwork (DIAN) to adaptively decouple invariant and variant patterns for enhancing MST forecasting.

Specifically, DIAN includes two modules: the Spatial Decoupled Invariant-Variant Learning module and the Temporal Augmented Invariant-Variant Learning module, designed to disentangle patterns in spatial and temporal dimensions, respectively. Drawing inspiration from [Bai *et al.*, 2020], the spatial module employs a spatial invariant-variant attention mechanism to distinguish between spatial invariant and variant patterns. After conducting decoupled invariant-variant convolution, the spatially-learned representations successfully captured both invariant and variant patterns. In the Temporal Augmented Invariant-Variant Learning module, we introduce the Temporal Invariant-Variant Attention to capture correlations between individual timestamps and the overarching temporal representation. To address pronounced temporal distribution shift, we incorporate the Temporal Intervention Mechanism, generating intervened samples for enhanced adaptability to potential shift. In addition, we employ a Joint Optimization strategy to reduce forecasting errors.

Our contributions can be summarized as:

- We introduce a unique perspective on invariant and variant patterns for MTS forecasting.
- We propose DIAN to address the invariant-variant pattern decoupling problem with novel invariant-variant attention designs for the spatial and temporal dimensions, respectively.
- We propose Temporal Intervention Mechanism to against potential distribution shift by reassembling variant patterns among timestamps.
- We conduct extensive experiments on five real-world dataset. The experiment results show consistent superiority in terms of effectiveness and efficiency compared with state-of-the-art methods.

2 Problem Formulation

We study the problem of multivariate time series forecasting with regard to historical time-evolving multiple variables. Given a time series dataset with N series and T timestamps, let $X_t = \{x_t^{(1)}, \cdots, x_t^{(i)}, \cdots, x_t^{(N)}\} \in R^N$ stands for multivariate values of N series where $x_t^{(i)} \in R^1$ represents the value of i-th series (variate) at timestamp t. Then, we define the historical observations of length L at timestamp t, $\mathbf{X}_{t-L+1:t} = \{X_{t-L+1}, \cdots, X_t\}$ as the lookback window, and the future values of length H, $\mathbf{X}_{t+1:t+H} = \{X_{t+1}, \cdots, X_{t+H}\}$ as the horizon window.

In order to model the relationships among series, we define a dynamic graph structure $\mathcal{G}_t = (\mathbf{X}_{t-L+1:t}, A_t)$, where $A_t \in R^{N\times N}$ is the adjacency matrix and $\mathbf{X}_{t-L+1:t} \in R^{N\times L}$ is the lookback window at timestamp t. Based on the above notations, we can formulate the *multivariate time series forecasting* task as:

$$\widehat{\mathbf{Y}}_{t+1:t+H} = \mathcal{F}_\Theta(\mathcal{G}_t) \tag{1}$$

where $\widehat{\mathbf{Y}}_{t+1:t+H} = \{\widehat{Y}_{t+1}, \cdots, \widehat{Y}_{t+H}\}$ are the predicted values corresponding for the horizon window; $\mathcal{F}_\Theta$ is a mapping function with parameters Θ; the dynamic graph structure $\mathcal{G}_t$ is learned during training.

3 Methodology

In this section, we illustrate the proposed Decoupled Invariant Attention Network (DIAN) for better modeling and decoupling invariant and variant patterns for enhancing multivariate time series forecasting performance. We first show an overview of our model, and then introduce each component of this model in detail.

3.1 Framework Overview

Figure 1 provides an overview of our framework, known as the Decoupled Invariant Attention Network (DIAN). In Section 3.2, to capture inter-series correlations and dynamic graph structures, we propose Spatial Decoupled Invariant-Variant Learning. This approach allows us to derive novel representations that incorporate relationships among different series. Moving on to Section 3.3, we propose Temporal Augmented Invariant-Variant Learning. This technique enables us to obtain representations that capture relationships among timestamps and generate multiple intervened samples, which simulate potential distribution shift. Finally, in Section 3.4, we optimize our model with Joint Optimization.

3.2 Spatial Decoupled Invariant-Variant Learning

Assumption 1

Given the dynamic graph structure $\mathcal{G}_t$ of time series, there exist spatial invariant patterns Q_I^t that reflect core spatial (inter-series) relationships, and spatial variant patterns Q_V^t that encompass a limited amount of valuable information while being intermingled with distribution shift. In this scenario, a function g is assumed to leverage spatial invariant-variant patterns and an adjacency matrix to acquire knowledge about the forthcoming dynamic graph structure at the subsequent timestamp, formally by $\mathcal{G}_{t+1} = g(Q_I^t, Q_V^t, A_t) + \epsilon_1$ and $Q_I^t = \mathbf{X}_{t-L+1:t}/Q_V^t$.

Spatial Invariant-Variant Attention

Based on the aforementioned assumption, it is understood that the time series can be decomposed into two components: spatial invariant and variant patterns. These patterns capture the essential relationships between the series, and it is necessary to design a mapping function, to extract the inter-series

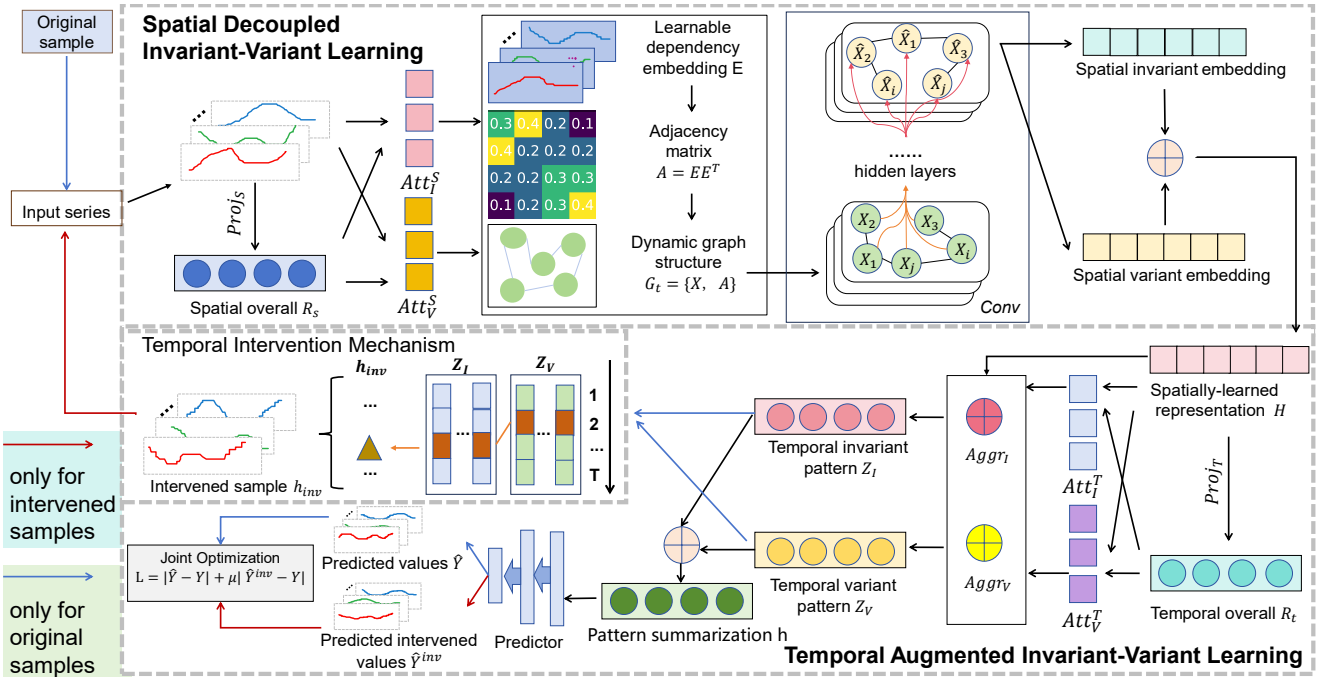


Figure 1: Framework Overview. This is an overview of DIAN, where each time series can have multiple features. However, for the sake of convenience, we use feature=1 as an example for plotting. In the graph, the blue line represents the unique data flow of the original time series, the red line represents the unique data flow of the intervened samples, and the black line represents the steps that both have undergone.

relationships from $\mathcal{G}$ and decouple these two patterns, facilitating subsequent predictions.

For this aim, we design a spatial invariant-variant attention network, to enable each series attending to the global inter-series representation for better understanding relationships among series. We capture meaningful spatial dependencies within the global inter-series structure by studying the contribution of each local series, thereby establishing a connection between the local and global contexts. Specifically, we first obtain a spatial overall representation of the current time series using a linear projection layer:

$$\mathcal{R}_s = Proj_S(\mathbf{X}_{t-L+1:t}), \tag{2}$$

where $\mathcal{R}_s \in R^{1\times T\times d_c}$ represents the spatial overall representation of the current inter-series relationships with d_c dimensions; $Proj_S$ represents a linear projection to map series dimension to one; $\mathbf{X}_{t-L+1:t}$ is a lookback window length time series. Then, we need to learn the dependencies between each series and the spatial overall representation. Because the learned representation represents the integrated spatial relationships of the entire time series, a better integration of the inter-series relationships can be achieved, by using the relationship between each individual series (local) and the spatial overall representation (global) as a guidance. For the i-th variate (series) $\mathbf{X}^{(i)}_{t-L+1:t}$ of the lookback window, we define the attention score as the spatial invariant attention score, which is given by:

$$\mathrm{Attn}^S_I(i) = \mathrm{SoftMax}(\frac{W_{qs}(\mathcal{R}_s)W_{ks}(\mathbf{X}^{(i)}_{t-L+1:t})}{\sqrt{d_S}}), \tag{3}$$

where $\mathrm{Attn}^S_I(i)$ represents the spatial invariant attention score of variate i; d_S represents the hidden dimension; W_{qs} and W_{ks} are the embedding layers for query vectors $\mathcal{R}_s$ and key vectors $\mathbf{X}^{(i)}_{t-L+1:t}$, respectively. In addition to using invariant attention to learn the core relationships, we also model the unstable inter-series patterns for spatial learning. Following [Zhang *et al.*, 2022], for each i-th variate $\mathbf{X}^{(i)}_{t-L+1:t}$ of the lookback window, we define the spatial variant attention score as:

$$\mathrm{Attn}^S_V(i) = \mathrm{SoftMax}(-\frac{W_{qs}(\mathcal{R}_s)W_{ks}(\mathbf{X}^{(i)}_{t-L+1:t})}{\sqrt{d_S}}), \tag{4}$$

where $\mathrm{Attn}^S_V(i)$ represents the spatial variant attention score of variate i. In this way, the spatial invariant and variant attention scores have a negative correlation, making it easy to distinguish core spatial relations and unstable spatial relations so that we can integrate them reasonably for forecasting.

Decoupled Invariant-Variant Convolution

Previous studies have demonstrated the capability of learning graph structures autonomously, bypassing the need for prior knowledge and thereby circumventing the issue of inaccurate input graphs [Bai *et al.*, 2020]. However, it is crucial to acknowledge that the previous studies also introduce the risk of learning inaccurate graphs. To tackle this challenge, we propose a decoupled invariant-variant convolution network for multivariate time series forecasting. First, we initialize learnable dependency embeddings $\mathbf{E} \in R^{N\times d_N}$ with d_N as the hidden dimension, and then the adjacency matrix A_t can be represented as:

$$A_t = \mathrm{SoftMax}(\mathrm{ReLU}(\mathbf{E}\mathbf{E}^T). \quad (5)$$

In order to enhance the reliability of the convolutional network utilized in graph learning, it is essential to incorporate a dependable relationship between series. Specifically, we incorporate spatial invariant and variant attention to capture the appropriate series relationships within the graph. By utilizing spatial invariant attention scores, we obtain an invariant embedding that represents the stable aspects of inter-series relationships. Similarly, spatial variant attention scores yield a variant embedding that captures the changeable components of these relationships. We formally define the ℓ-th layer of such networks by:

$$\mathbf{H}_I^{\ell+1} = f(\mathbf{H}^\ell, A_t, \mathrm{Attn}_I^S) = \sigma(\mathrm{Attn}_I^S \mathbf{H}^\ell A_t \mathcal{W}_I^{\mathbf{H}}), \quad (6)$$

$$\mathbf{H}_V^{\ell+1} = f(\mathbf{H}^\ell, A_t, \mathrm{Attn}_V^S) = \sigma(\mathrm{Attn}_V^S \mathbf{H}^\ell A_t \mathcal{W}_V^{\mathbf{H}}), \quad (7)$$

where $\mathbf{H}^\ell \in R^{N\times T\times d_c}$ represents the hidden representation with d_c dimensions of ℓ-th layer; $\mathbf{H}_I^{\ell+1}$ and $\mathbf{H}_V^{\ell+1}$ represent the spatial invariant and variant embeddings respectively. $\mathcal{W}_I^{\mathbf{H}}$ and $\mathcal{W}_V^{\mathbf{H}}$ represents weights in invariant and variant embedding learning; σ is an activation function. Note that when $\ell = 0$, we have $\mathbf{H}^0 = \mathbf{X}_{t-L+1:t}$. Then, the learned representation incorporating both invariant and invariant patterns can be represented as

$$\mathbf{H}^{\ell+1} = \mathbf{H}_I^{\ell+1} + \alpha \mathbf{H}_V^{\ell+1}, \quad (8)$$

where α is a hyperparameter to balance invariant and variant embeddings.

3.3 Temporal Augmented Invariant-Variant Learning

Assumption 2

Given the multivariate window at timestamp t, $\mathbf{X}_{t-L+1:t}$, there exist temporal invariant patterns P_I^t that usually reflect core temporal relations, and temporal variant patterns P_V^t that represent temporal information mixed with distribution shift. Assume there exists a function y that can project temporal invariant and variant patterns into future values, written as: $\mathbf{X}_{t+1:t+H} = y(P_I^t, P_V^t) + \epsilon_2$ and $P_I^t = \mathbf{X}_{t-L+1:t}/P_V^t$.

Temporal Invariant-Variant Attention

In order to model temporal dependencies, following Assumption 2, we aim to design a function to effectively decompose the spatially-learned representation of time series, $\mathbf{H}^{\ell+1}$ into temporal invariant and variant patterns. To achieve this goal, we propose temporal invariant-variant attention mechanism to model temporal dependencies. Specially, we first obtain an overall temporal representation of $\mathbf{H}^{\ell+1}$ using a linear projection:

$$\mathcal{R}_T = \mathit{Proj}_T(\mathbf{H}^{\ell+1}), \quad (9)$$

where $\mathcal{R}_T \in R^{N\times 1\times d_c}$ represents the temporal overall representation of $\mathbf{H}^{\ell+1}$ with d_c dimensions; Proj_T represents a linear projection to map time dimension to one; Then, we learn the invariant-variant attention between each timestamp and the temporal overall representation. The temporal overall representation encompasses the temporal information, such as seasonal and trend, within a lookback window. This approach enables us to understand the contribution of each timestamp to the temporal overall representation, and establishes the connection between the local and global aspects. Specifically, for each t-th timestamp, we define the temporal invariant and variant attention score as:

$$\mathrm{Attn}_I^T(t) = \mathrm{SoftMax}(\frac{W_{qt}(\mathcal{R}_T)W_{kt}(\mathbf{H}_t^{\ell+1})}{\sqrt{d_M}}), \quad (10)$$

$$\mathrm{Attn}_V^T(t) = \mathrm{SoftMax}(-\frac{W_{qt}(\mathcal{R}_T)W_{kt}(\mathbf{H}_t^{\ell+1}))}{\sqrt{d_M}}), \quad (11)$$

where $\mathrm{Attn}_I^T(t)$ and $\mathrm{Attn}_V^T(t)$ represent the temporal invariant and variant attention score respectively; $\mathbf{H}_t^{\ell+1}$ represents the spatially-learned representation at timestamp t; d_M represents the hidden dimension; W_{qt} and W_{kt} represent the embedding layers for query vectors $\mathcal{R}_T$ and key vectors $\mathbf{H}_t^{\ell+1}$ from the time dimension. Then, for each $\mathbf{H}_t^{\ell+1}$ at t-th timestamp, we want to aggregate the patterns for representation:

$$\mathbf{z}_I^t = \mathrm{Aggr}_I(\mathbf{H}_t^{\ell+1}, \mathrm{Attn}_I^T(t)), \quad (12)$$

$$\mathbf{z}_V^t = \mathrm{Aggr}_V(\mathbf{H}_t^{\ell+1}, \mathrm{Attn}_V^T(t)), \quad (13)$$

where Aggr_I and Aggr_V are aggregation functions for temporal invariant and variant patterns respectively; $\mathbf{z}_I^t$ and $\mathbf{z}_V^t$ represents the pattern aggregations of temporal invariant and variant patterns. Then, for timestamp t, we acquire pattern summarization $\mathbf{h}^t = \mathbf{z}_I^t + \beta * \mathbf{z}_V^t$ fed into subsequent layers, where β is a hyperparameter to balance the temporal invariant and variant patterns.

Temporal Intervention Mechanism

Distribution shift in time series pose a significant challenge to the accuracy of time series forecasting, as they often exhibit temporal fluctuations that can seriously impede prediction performance. Additional analysis on distribution shift in time series is provided in Appendix[1]. Therefore, we further introduce a temporal intervention design into the temporal learning. Specially, we first denote the timestamps in a lookback window of length L as $\{t_1, t_2, \cdots, t_L\}$. Then, we let the intervention process accomplished by randomly replacing temporal variant patterns. Formerly, for $i, j \in [1, L]$, we generate an intervened representation:

$$s_I^{t_i} = z_I^{t_i}, s_V^{t_i} = z_V^{t_j}, \quad (14)$$

where $s_I^{t_i}$ and $s_V^{t_i}$ represent the intervened invariant and variant patterns respectively. Then, the intervened invariant and variant patterns are added up as intervened representation as: $\mathbf{h}_{inv}^{t_i} = s_I^{t_i} + \gamma s_V^{t_i}$, where $\mathbf{h}_{inv}^t$ is the intervened sample at t-th timestamp and γ is a hyperparameter to balance the intervened invariant and variant patterns for forecasting.

[1]https://github.com/xhh39/DIAN

3.4 Joint Optimization

In time series forecasting, the classic objective function is set as the Mean Absolute Error,

$$\mathcal{L} = |\hat{\mathbf{Y}}_{t+1:t+H} - \mathbf{X}_{t+1:t+H}|, \tag{15}$$

Idealy, if the essential information is totally extracted and there is no distribution shift mixed with the information, we only use invariant patterns to make prediction as: $\hat{\mathbf{Y}}_{t+1:t+H} = f(P_I^t, R_I^t)$, where f is a fully connected layer predictor. In order to adapt to the distribution shift that often occur in time series forecasting tasks, we introduce both invariant and variant patterns to make prediction as: $\hat{\mathbf{Y}}_{t+1:t+H} = f(P_I^t, R_I^t, P_V^t, R_V^t)$. In section 3.3, we propose the temporal intervention mechanism and obtain intervened samples. In order to simulate various distribution shift that may occur and further adapt to distribution shift, we use the obtained intervened samples to make prediction as: $\hat{\mathbf{Y}}^{inv}_{t+1:t+H} = f(s_I^t, s_V^t)$. Then, new joint optimization with regard to temporal intervention mechanism can be formalized as:

$$\begin{aligned} \min \mathcal{L}_{final} = \min &|\hat{\mathbf{Y}}_{t+1:t+H} - \mathbf{X}_{t+1:t+H}| \\ &+ \lambda|\hat{\mathbf{Y}}^{inv}_{t+1:t+H} - \mathbf{X}_{t+1:t+H}|, \end{aligned} \tag{16}$$

where $\mathcal{L}_{final}$ is minimized to exploit invariant and variant patterns while discovering and adapting to the distribution shift ahead of time; λ is a hyperparameter to balance between two objectives.

4 Experiments

We conduct extensive experiments on five real-world time series benchmarks and compare our model with many effective time series forecasting models (including state-of-the-art graph neural network-based models) to validate the performance of our model.

4.1 Experimental Setup

Datasets

We evaluate our proposed method on five real-world datasets and use the min-max normalization to normalize all these datasets. Except for the COVID-19 dataset, we split the datasets into training, validation, and test sets with the ratio of 7:2:1 in a chronological order. For the COVID-19 dataset, the ratio is 6:2:2 because of the limitation of data scale in temporal dimension. More detailed information about the datasets is provided in Appendix[1].

Evaluation

To compare the performance of different forecasting models, we deploy two widely used evaluation metrics: 1) Root Mean square Error. 2) Mean Absolute Error. More detailed information about the evaluation is provided in Appendix[1].

Implementation

We use PyTorch to implement our model and baselines. All models were evaluated on a Linux server with one RTX 3090 GPU. We use MAE (Mean Absolute Errors) as the loss function and the Adam Optimizer with a learning rate of 1e-3 with proper early stopping. For the main experiment, we fix the lookback window length as 12 and the horizon window length as 12. More detailed information about the implementation is provided in Appendix[1].

Baselines

To verify the effectiveness of our model, we compared it with several representative baseline methods on the five datasets. The baseline methods mainly include: (1) *classic method VAR* [Watson, 1993]; (2) *deep learning-based models* such as SFM [Zhang *et al.*, 2017], LSTNet [Lai *et al.*, 2018], TCN [Bai *et al.*, 2018], DeepGLO [Sen *et al.*, 2019], and CoST [Woo *et al.*, 2022b]; GNN-based models such as GraphWaveNet [Wu *et al.*, 2019], StemGNN [Cao *et al.*, 2020b], MTGNN [Wu *et al.*, 2020], and AGCRN [Bai *et al.*, 2020]; Transformer-based models including Reformer [Kitaev *et al.*, 2020], Informer [Zhou *et al.*, 2021], Autoformer [Wu *et al.*, 2022] and FEDformer [Zhou *et al.*, 2022b]. In addition, we also compare DIAN with SOTA TAMP-S2GCNets [Chen *et al.*, 2021]. To compare fairly, our experiments for baselines and our model are under the same experimental settings. Due to spatial constraints, more baseline implementation details can be found in Appendix[1].

4.2 Overall Performance

Table 1 demonstrates the overall performance of thirteen baselines and DIAN. It is obvious that DIAN achieves a new state-of-the-art on all datasets. Compared with the best-performing across all datasets, DIAN makes an improvement of 11.3% in MAE and 7.9% in RMSE. Notably, we find that on COVID-19 dataset, transformer-based models like Reformer achieve a competitive performance because they excel at capturing temporal dependencies of time series. However, they are not as good as the GNN-based models at capturing spatial dependencies. As a result, on Wiki, Traffic, and ECG datasets, GNN-based models like AGCRN and MTGNN achieve a more promising performance. DIAN not only learns spatial and temporal dependencies simultaneously, but also integrates invariant and variant patterns reasonably, therefore, it outperforms the baseline models.

4.3 Parameters and Model Analysis

Efficiency Analysis

In order to verify the high efficiency of DIAN, we investigate the parameter volumes and training time costs of DIAN and some lightweight baseline models (StemGNN, AGCRN, GraphWaveNet, MTGNN) on Traffic and Wiki datasets. Table 2 shows the comparison of parameter volumes and average training time costs over five rounds of experiments. Obviously, we can find that DIAN always have the lowest volume of parameters among the comparative models. Compared with the baseline models, DIAN achieves a reduction of more than 70.7% and 50.2% in parameter volumes on Traffic and Wiki datasets respectively. In addition, DIAN runs much faster than other baselines by at least 56.3% and 46.9% on Traffic and Wiki datasets respectively. The reason is that DIAN captures relationships between each series/timestamp and the overall representation instead of learning relationships among timestamps/series.

Models	Traffic		ECG		COVID-19		Wiki		Solar	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
VAR	0.535	1.133	0.120	0.170	0.226	0.326	0.057	0.094	0.184	0.234
SFM	0.029	0.044	0.095	0.135	0.205	0.308	0.081	0.156	0.161	0.283
LSTNet	0.026	0.057	0.079	0.115	0.248	0.305	0.054	0.090	0.148	0.200
TCN	0.052	0.067	0.078	0.107	0.317	0.354	0.094	0.142	0.176	0.222
DeepGLO	0.025	0.037	0.110	0.163	0.169	0.253	0.110	0.113	0.178	0.400
Reformer	0.029	0.042	0.062	0.090	0.152	0.209	0.048	0.085	0.234	0.292
Informer	0.020	0.033	0.056	0.085	0.200	0.259	0.051	0.086	0.151	0.199
Autoformer	0.029	0.043	0.055	0.081	0.159	0.211	0.069	0.103	0.150	0.193
FEDformer	0.025	0.038	0.055	0.080	0.160	0.219	0.068	0.098	0.139	0.182
GraphWaveNet	0.013	0.034	0.093	0.142	0.201	0.255	0.061	0.105	0.183	0.238
StemGNN	0.080	0.135	0.100	0.130	0.421	0.508	0.190	0.255	0.176	0.222
MTGNN	0.013	0.030	0.090	0.139	0.394	0.488	0.101	0.140	0.151	0.207
AGCRN	0.084	0.166	0.055	0.080	0.254	0.309	0.044	0.079	0.123	0.214
DIAN (ours)	**0.013**	**0.029**	**0.049**	**0.075**	**0.128**	**0.175**	**0.040**	**0.074**	**0.095**	**0.167**

Table 1: Overall performance comparisons of forecasting models on the five datasets.

Models	Traffic		Wiki	
	Parameters	Training(s/epoch)	Parameters	Training(s/epoch)
GraphWaveNet	280, 860	81.33 $\pm$1.22	292, 460	15.87 $\pm$0.35
StemGNN	1, 606, 140	178.56$\pm$1.92	4, 102, 406	84.56 $\pm$1.13
AGCRN	749, 940	108.81$\pm$1.34	755, 740	19.86$\pm$0.97
MTGNN	707, 516	154.16 $\pm$1.09	1, 533, 436	25.41 $\pm$0.57
DIAN(ours)	**82263**	**35.53**$\pm$1.17	**145,653**	**8.43**$\pm$0.18

Table 2: Comparisons of parameter volumes and training time costs on datasets Traffic and Wiki.

Ablation Study

In this section, we conduct ablation studies to verify the effectiveness of the proposed Spatial Decoupled Invariant-Variant Learning and Temporal Augmented Invariant-Variant Learning in DIAN on Solar and COVID-19 datasets. Specifically, w/o spatial is DIAN without Spatial Decoupled Invariant-Variant Learning and w/o temporal represents DIAN without Temporal Augmented Invariant-Variant Learning. We verify the necessity of capturing inter-series (spatial) correlations and intra-series (temporal) dependencies by removing these two parts respectively. The results presented in Table 3 demonstrate that the removal of either spatial or temporal components leads to a decline in model performance, underscoring the effectiveness of each component.

4.4 Visualization

To gain a better understanding of DIAN which can distinguish between invariant patterns and variant patterns in both temporal and spatial dimensions in multivariate time series forecasting. We conduct visualization experiments on the COVID-19 dataset.

Visualization of the Spatial Invariant and Variant Attention Score Learned by DIAN

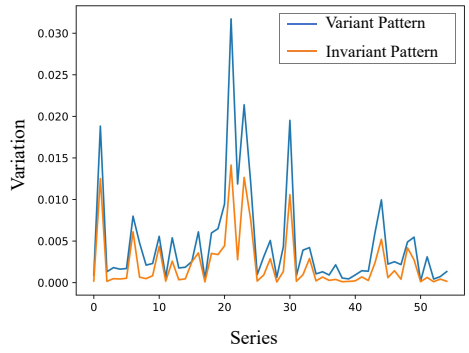
To demonstrate the capability of Spatial Decoupled Invariant-Variant Learning in capturing both stable and potentially volatile spatial relationships, we employed the COVID-19 dataset and visualized the variations of the invariant attention scores and variant attention scores using line graphs within the same lookback window but across different timestamps. Specifically, we computed the differences between each timestamp's attention score and the previous timestamp's attention score (excluding the first timestamp), resulting in $(L-1)$ difference values. Taking the average of these difference values provided a measure of the intensity of spatial relationship changes within the lookback window for each series. The results are illustrated in Figure 2b, where the blue line represents the variation of the variant attention score, and the orange line represents the variation of the invariant attention score. From the results, it is evident that the variant attention score exhibits more pronounced changes compared to the invariant attention score across all 55 series. Therefore, we can conclude that Spatial Decoupled Invariant-Variant Learning can effectively capture and distinguish stable spatial relationships from potentially unstable ones.

Visualization of Temporal Invariant and Variant Pattern Learned by DIAN

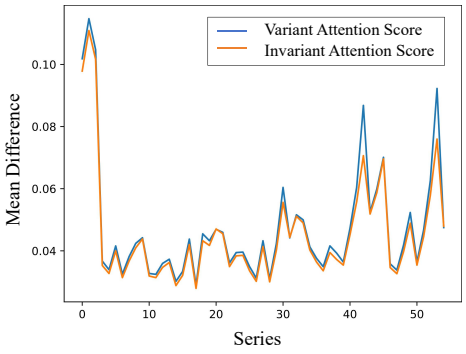
To validate the conformity of temporal invariant and variant patterns with our proposed viewpoint, which states that the variation amplitude of invariant patterns is smaller and more regular, while variant patterns exhibit drastic changes, we utilized the COVID-19 dataset as an example. We visualized the variance of each series within a lookback window using a line

Metrics	Solar			COVID-19		
	w/o spatial	w/o temporal	DIAN	w/o spatial	w/o temporal	DIAN
MAE	0.117	0.097	**0.095**	0.132	0.158	**0.128**
RMSE	0.200	0.172	**0.167**	0.175	0.205	**0.175**

Table 3: Ablation study on Solar and COVID-19 dataset.



(a) Temporal



(b) Spatial

Figure 2: Visualization results on the COVID-19 dataset.

graph, as shown in Figure 2a. The blue line represents the variance of variant patterns within a lookback window, while the orange line represents the variance of invariant patterns within the same window. The x-axis denotes the series number. From the results, it is evident that the variance of variant patterns consistently exceeds that of invariant patterns. Thus, we can conclude that through the Temporal Invariant-Variant Attention mechanism, we have successfully decomposed the series into invariant and variant components. The invariant pattern captures the relatively stable part, while the variant pattern captures the part that exhibits significant changes with variations in the environment.

5 Related Work

5.1 Time Series Forecasting

Time series forecasting (TSF) has gained significant attention due to its practical significance. In recent years, notable advancements have been made in time series forecasting research. DeepAR [Flunkert *et al.*, 2020] employs an RNN structure for accurate predictions. Nbeats [Oreshkin *et al.*, 2019] achieved remarkable improvements by utilizing residual computation and multiple fully connected layers, producing interpretable outputs. Moreover, transformer-based models have gained popularity in the field of time series forecasting, such as Informer [Zhou *et al.*, 2021] and Autoformer [Wu *et al.*, 2022]. In addition, there have been some recent works that have achieved new heights in the field of time series forecasting [Yi *et al.*, 2024b; Yi *et al.*, 2024a; Hu *et al.*, 2023; Ren *et al.*, 2022; Zhang *et al.*, 2024].

5.2 Graph Convolutional Networks Based Forecasting Model

GCN, a type of GNN specialized for graph-structured data, has wide applications in node classification, link prediction, and graph classification [Wu *et al.*, 2021]. It extracts relationship features between nodes, yielding satisfactory results. Recently, graph learning has been employed in time series forecasting to extract relationships in node or series data. However, traditional GCN relies on a predefined graph as input, which is difficult to obtain and less suitable for dynamic data like traffic flow [Bai *et al.*, 2020]. AGCRN addresses this challenge by capturing node-specific patterns and automatically inferring inter-dependencies among traffic series, which has demonstrated excellent performance in traffic forecasting [Bai *et al.*, 2020]. Nevertheless, the stability of the adjacency matrix learning in AGCRN can be improved further. Research on temporal graphs has been rapidly advancing recently, and many methods are worth considering and drawing inspiration from [Dong *et al.*, 2024; Dong *et al.*, 2023].

6 Conclusion Remarks

In this paper, we propose a multivariate time series forecasting model, Decoupled Invariant Attention Network (DIAN), to decouple invariant and variant patterns in both spatial and temporal dimensions and combine them reasonably for forecasting. First, we use Spatial Invariant-Variant Attention to calculate spatial invariant and variant attention scores for capturing inter-series correlations and propose Decoupled Invariant-Variant Convolution to get spatially-learned representations. Second, we propose Temporal Invariant-Variant Attention to capture temporal correlations and Temporal Intervention Mechanism to create intervened samples to simulate potential distribution shift. Extensive experiments on five real-world datasets demonstrate that our model can achieves state-of-the-art performances with higher efficiency and fewer parameters.

Acknowledgments

This research is funded by the Science and Technology Development Fund (FDCT), Macau SAR (file no. 0123/2023/RIA2, 001/2024/SKL), the Start-up Research Grant of University of Macau (File no. SRG2021-00017-IOTSC).

Contribution Statement

In this work, Haihua Xu and Wei Fan have an equal contribution, specifically, they led the project, provided theoretical support, and were responsible for the overall model design, code implementation, experimental design and paper writing. Kun Yi provided guidance in solving complex problems. Pengyang Wang provided valuable feedback on the paper drafts. All authors reviewed and approved the final manuscript.

References

[Bai *et al.*, 2018] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018.

[Bai *et al.*, 2020] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. 2020.

[Beccali *et al.*, 2008] M Beccali, M Cellura, V Lo Brano, and Antonino Marvuglia. Short-term prediction of household electricity consumption: Assessing weather sensitivity in a mediterranean area. *Renewable and Sustainable Energy Reviews*, 12(8):2040–2065, 2008.

[Cao *et al.*, 2020a] Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, et al. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in neural information processing systems*, 33:17766–17778, 2020.

[Cao *et al.*, 2020b] Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, et al. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in neural information processing systems*, 33:17766–17778, 2020.

[Chen *et al.*, 2021] Yuzhou Chen, Ignacio Segovia-Dominguez, Baris Coskunuzer, and Yulia Gel. Tamps2gcnets: coupling time-aware multipersistence knowledge representation with spatio-supra graph convolutional networks for time-series forecasting. In *International Conference on Learning Representations*, 2021.

[Dong *et al.*, 2023] Hao Dong, Zhiyuan Ning, Pengyang Wang, Ziyue Qiao, Pengfei Wang, Yuanchun Zhou, and Yanjie Fu. Adaptive path-memory network for temporal knowledge graph reasoning. *arXiv preprint arXiv:2304.12604*, 2023.

[Dong *et al.*, 2024] Hao Dong, Pengyang Wang, Meng Xiao, Zhiyuan Ning, Pengfei Wang, and Yuanchun Zhou. Temporal inductive path neural network for temporal knowledge graph reasoning. *Artificial Intelligence*, 329:104085, 2024.

[Fan *et al.*, 2022] Wei Fan, Shun Zheng, Xiaohan Yi, Wei Cao, Yanjie Fu, Jiang Bian, and Tie-Yan Liu. Depts: deep expansion learning for periodic time series forecasting. *arXiv preprint arXiv:2203.07681*, 2022.

[Fan *et al.*, 2023] Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. Dish-ts: A general paradigm for alleviating distribution shift in time series forecasting, 2023.

[Flunkert *et al.*, 2020] Valentin Flunkert, David Salinas, and Jan Gasthaus. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3), 2020.

[Hu *et al.*, 2023] Xuanming Hu, Wei Fan, Kun Yi, Pengfei Wang, Yuanbo Xu, Yanjie Fu, and Pengyang Wang. Boosting urban prediction via addressing spatial-temporal distribution shift. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 160–169. IEEE, 2023.

[Jiang *et al.*, 2022] Song Jiang, Tahin Syed, Xuan Zhu, Joshua Levy, Boris Aronchik, and Yizhou Sun. Bridging self-attention and time series decomposition for periodic forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3202–3211, 2022.

[Kitaev *et al.*, 2020] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

[Lai *et al.*, 2018] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks, 2018.

[Lim and Zohren, 2021] Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021.

[Oreshkin *et al.*, 2019] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. 2019.

[Ren *et al.*, 2022] Weijieying Ren, Pengyang Wang, Xiaolin Li, Charles E Hughes, and Yanjie Fu. Semi-supervised drifted stream learning with short lookback. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1504–1513, 2022.

[Sen *et al.*, 2019] Rajat Sen, Hsiang-Fu Yu, and Inderjit S Dhillon. Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[Wang *et al.*, 2022] Zhiyuan Wang, Xovee Xu, Weifeng Zhang, Goce Trajcevski, Ting Zhong, and Fan Zhou. Learning latent seasonal-trend representations for time series forecasting. *Advances in Neural Information Processing Systems*, 35:38775–38787, 2022.

[Watson, 1993] Mark W. Watson. Vector autoregressions and cointegration. *Working Paper Series, Macroeconomic Issues*, 4, 1993.

[Woo *et al.*, 2022a] Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Cost: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. *arXiv preprint arXiv:2202.01575*, 2022.

[Woo *et al.*, 2022b] Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Cost: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. *arXiv preprint arXiv:2202.01575*, 2022.

[Wu *et al.*, 2019] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1907–1913. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[Wu *et al.*, 2020] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 753–763, 2020.

[Wu *et al.*, 2021] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, (1):32, 2021.

[Wu *et al.*, 2022] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting, 2022.

[Yi *et al.*, 2024a] Kun Yi, Qi Zhang, Wei Fan, Hui He, Liang Hu, Pengyang Wang, Ning An, Longbing Cao, and Zhendong Niu. Fouriergnn: Rethinking multivariate time series forecasting from a pure graph perspective. *Advances in Neural Information Processing Systems*, 36, 2024.

[Yi *et al.*, 2024b] Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain mlps are more effective learners in time series forecasting. *Advances in Neural Information Processing Systems*, 36, 2024.

[Yu *et al.*, 2017] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

[Zeng *et al.*, 2023] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? 2023.

[Zhang *et al.*, 2017] Liheng Zhang, Charu Aggarwal, and Guo-Jun Qi. Stock price prediction via discovering multi-frequency trading patterns. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2141–2149, 2017.

[Zhang *et al.*, 2022] Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, Zhou Qin, and Wenwu Zhu. Dynamic graph neural networks under spatio-temporal distribution shift. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 6074–6089. Curran Associates, Inc., 2022.

[Zhang *et al.*, 2024] Zhaofan Zhang, Yanan Xiao, Lu Jiang, Dingqi Yang, Minghao Yin, and Pengyang Wang. Spatial-temporal interplay in human mobility: A hierarchical reinforcement learning approach with hypergraph representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9396–9404, 2024.

[Zheng *et al.*, 2015] 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.

[Zhou *et al.*, 2021] 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, 2021.

[Zhou *et al.*, 2022a] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Zhou *et al.*, 2022b] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. 2022.