MUSE-Net: Disentangling Multi-Periodicity for Traffic Flow Forecasting

Jianyang Qin[1], Yan Jia[1,2], Yongxin Tong[3], Heyan Chai[1], Ye Ding[4], Xuan Wang[1], Binxing Fang[1,2] and Qing Liao[1,2]✉
[1]Harbin Institute of Technology (Shenzhen), Shenzhen, China [2]Peng Cheng Laboratory, Shenzhen, China
[3]Beihang University, Beijing, China [4]Dongguan University of Technology, Dongguan, China
Email: 22b351005@stu.hit.edu.cn; jiayanjy@vip.sina.com; yxtong@buaa.edu.cn; chaiheyan@stu.hit.edu.cn; dingye@dgut.edu.cn; wangxuan@cs.hitsz.edu.cn; fangbx@cae.cn; liaoqing@hit.edu.cn

2024 IEEE 40th International Conference on Data Engineering (ICDE) | 979-8-3503-1715-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDE60146.2024.00103

***Abstract*—Accurate forecasting of traffic flow plays a crucial role in building smart cities in the new era. Previous work has achieved success in learning inherent spatial and temporal patterns of traffic flow. However, existing works investigated the multiple periodicities (e.g., hourly, daily, and weekly) of traffic via entanglement learning, which has not yet dealt with distribution shift and interaction shift problems in traffic flow. In this paper, we propose a novel disentanglement learning network, called MUSE-Net, to tackle the limitations of entanglement learning by simultaneously factorizing the exclusiveness and interaction of multi-periodic patterns in traffic flow. Grounded in the theory of mutual information, we first learn and disentangle exclusive and interactive representations of traffics from multi-periodic patterns. Then, we utilize semantic-pushing and semantic-pulling regularizations to encourage the learned representations to be independent and informative. Moreover, we derive a lower bound estimator to tractably optimize the disentanglement problem with multiple variables and propose a joint training model for traffic forecasting. Extensive experimental results on several real-world traffic datasets demonstrate the effectiveness of the proposed framework. The code is available at: https://github.com/JianyangQin/MUSE-Net.**

***Index Terms*—Traffic Flow Forecasting, Time Series, Multivariate, Disentanglement**

I. Introduction

With the rapid development of cities, the urban population is increasing, resulting in a growing number of instances of traffic congestion encountered by people in their daily commutes. To address traffic congestion, many countries are committed to vigorously developing the Intelligent Transportation System (ITS) [1]. Moreover, ITS is of great importance for many real-world applications, such as public safety and disaster control [2]. Traffic flow forecasting has played a critical role in ITS. Accurate traffic flow prediction can help the transportation department design better transportation scheduling and mobility management strategies. In general, the goal of traffic flow forecasting is to predict the traffic volume (e.g., inflow and outflow) of each region from historical traffic data [3], [4].

✉ Corresponding authors.

This work was partially supported by the National Key-Research and Development Program of China (Grant No.2020YFB2104003) and the National Natural Science Foundation of China (Grant No.U19A2067).

Some early studies have been proposed to address traffic flow forecasting by simply considering temporal and spatial data. These methods either only capture temporal correlation via statistics model [5], [6], Recurrent Neural Networks (RNN) [7]–[9] and Transformer [10], or combine spatial learning with temporal learning by further introducing Convolutional Neural Network (CNN) [11], [12] and Graph Neural Network (GNN) [13], [14] to learn grid-based or unstructured spatial dependency. More recently, some methods propose to capture the fine-grained temporal information to enrich the temporal representation by modeling the multiple periodicities in different time resolutions (e.g., hourly, daily, and weekly) [15]. Specifically, a sequence of traffic can be intercepted into closeness, period, and trend sub-series, which corresponds to hourly, daily, and weekly resolutions.

However, existing works learn this multi-periodicity in an entangled manner. One entanglement learning jointly encodes the multi-periodic sub-series into a unified representation, ignoring the difference among multi-periodicity [16], [17]. Another entanglement learning simply separates multi-periodic sub-series encoding without considering the similarity among multi-periodicity [18]–[20] . Therefore, how to decouple the similarity and difference existing in multi-periodic representations still face many challenges. The main challenges in modeling the multi-periodicity are as follows:

(1) Distribution shift. In the real world, a lot of external factors (e.g., weather, holidays, and traffic accidents) affect traffic and cause the traffic flow to change. Thus, the distribution of a time series may shift [21]. Fig. 1 illustrates two typical cases of the distribution shift problem, i.e., level shift and point shift. If we jointly learn an entangled representation from all multi-periodic sub-series, it would be challenging to model these distribution shifts existing in different sub-series. To address this, we propose to disentangle the multi-periodicity into several independent exclusive representations; that is, we use different networks to model the multiple sub-series in different time resolutions and maintain the differences among multi-periodicity, so as to better characterize each multi-periodic sub-series for traffic flow forecasting.

(2) Interaction shift. The interaction means that the ob-

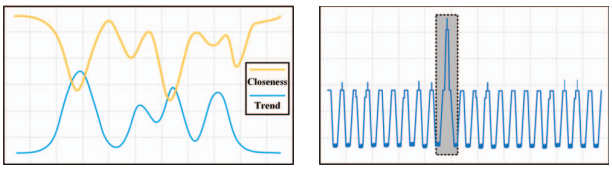
2375-026X/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDE60146.2024.00103

served time series may affect the forecasting of future traffic flow. In particular, future traffic flow may interact differently with multi-periodic sub-series, and each interaction may change over time, which is called interaction shift [16]. Fig. 2 illustrates an example of interaction shift problem. The reason for this interaction shift to appear is the semantic divergence among multiple sub-series. Particularly, the closeness sub-series characterizes short-term dependency, while the trend sub-series characterizes long-term dependency. Thus, we propose to learn an interactive representation to capture the common pattern of traffic flow sharing information among multi-periodicity, which can alleviate the interactive gaps between future traffic flow and multi-periodic sub-series.

(3) Optimizing multivariate disentanglement. In our case, we intercept a sequence of traffic flow into multiple sub-series, which can be seen as multiple variates. Recently, a variety of works [22] have been proposed to disentangle independent bivariates. However, as the number of variates increases, it is becoming harder for disentanglement learning to address increasing unknown posterior distributions and more complex relations between variates. Therefore, how to optimize multivariate disentanglement with independence and informativeness remains an open issue.

In this paper, we propose a novel predictive network, namely, **MU**lti-periodicity di**SE**ntanglement Network (**MUSE-Net**), to mitigate the limitation of entangled traffic flow forecasting by explicitly learning the disentangled multi-periodic patterns. In particular, we disentangle the traffic flow of the closeness, period, and trend sub-series into three exclusive representations with temporal peculiarity to alleviate the distribution shift problem, as well as an interactive representation shared across all time sub-series to tackle the interaction shift problem. Moreover, we introduce two regularization terms, i.e., semantic-pushing and semantic-pulling terms. The semantic-pushing term forces the interactive representation to be pushed away from arbitrary exclusive representations, ensuring the interactive representation to be independent of each exclusive representation. The semantic-pulling term forces the interactive representation to be pulled towards original closeness, period, and trend time sub-series, encouraging interactive representation to learn common patterns shared across different time sub-series. After that, the learned exclusive and interactive representations are aggregated to further capture spatial dependency for traffic flow forecasting. Finally, we propose a lower bound estimator to tackle the intractable disentanglement problem by optimizing the mutual information of exclusive and interactive representations. The main contributions of this paper can be summarized as follows:

- MUSE-Net proposes a multivariate disentanglement network to accurately model the multi-periodic patterns by decoupling exclusive and interactive representations, which can deal with the distribution shift and interaction shift for traffic flow forecasting.
- We introduce semantic-pushing and semantic-pulling regularization terms to encourage exclusive and interactive representations to be independent and informative.



(a) level shift (b) point shift

Fig. 1. Two typical cases of the distribution shift of time series. The level shift means that a time series e.g., closeness) is totally different from others (e.g., trend) in terms of distribution, while the point shift means that a time series includes outliers.

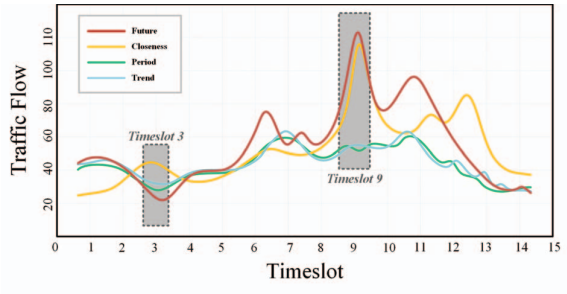


Fig. 2. An example to illustrate the interaction shift. Specifically, we sample the future traffic flow (i.e., traffic flow from t step to $t+15$ step) and closeness, period, and trend traffic flow related to the future traffic flow in hourly, daily and weekly dimensions, respectively. Then, we plot these traffic flow in the same timeslot axis to illustrate the correlations among them. At timeslot 3, the future traffic flow positively interacts with (i.e., is similar to) period and trend time sub-series, while negatively interacts with (i.e., is different from) the closeness time sub-series. As time goes by, the future traffic flow becomes close to the closeness time sub-series at timeslot 9.

- We drive a lower bound estimator to straightforwardly differentiate and optimize the problem of disentangled representation learning with multiple variates.
- Extensive experimental results on three traffic datasets demonstrate the superiority of the proposed method compared to state-of-the-art traffic forecasting methods.

II. Related Work

In this section, we briefly review the related work on traffic flow forecasting and disentanglement learning.

A. Traffic Forecasting

With the rapid development of the city, traffic forecasting, which models the changes of traffic conditions over time and across regions, has attracted increasing research attention. As representative methods, recurrent networks, such as Long-Short-Term-Memory (LSTM) networks [8] and Gated Recurrent Unit (GRU) networks [9], [23], learn temporal correlation from long-range sequences. For spatial learning, Convolution Neural Network (CNN) has been widely used in traffic forecasting of grid-cell data [15], [24]. Compared to the CNN-based method, Graph Neural Network (GNN) based methods, such as [19], [25]–[35], were generalized to excavate the spatial dependency of nonlinear structured data. Specifically, Data-Driven Spatial-Temporal Graph Neural Network (STGNN-DJD) [14] developed two novel graphs

to model the flow characteristic and pattern correlation, respectively. Furthermore, several studies [36]–[38] introduced attention mechanisms to better learn contextual information from traffic by adaptively focusing on the most relevant features to predictive data [16]. For example, Spatio-Temporal Wavelets (STWave) [39] modeled trends and events through a disentangled dual-channel network and then captured dynamic spatial correlations through the graph attention mechanism.

Although these methods have remarkably improved the performance of traffic prediction, they are ways of entangled learning that lack thoughtful consideration of the multi-periodic patterns for traffic forecasting. For example, both [16] and [17] are difficult to capture intrinsic patterns in different temporal dimensions by learning a unified representation of multi-periodicity. Although [18] and [20] separate the encoding of multi-periodicity and fuse multi-periodicity by gating mechanism and convolution, respectively, they hardly distinguish the similarity and difference among multi-periodicity and thus learn redundant information. By contrast, our proposed method explicitly decouples the entangled time sub-series into exclusive and interactive representations, which not only captures the private patterns of each time sub-series but also captures the common patterns among multiple sub-series. In this way, our proposed method can be powerful in overcoming the distribution shift and interaction shift problems.

B. Disentanglement Learning

Disentanglement learning [21], [40] is to factorize the observed data into several different representations that characterize the underlying explanatory factors. Variational Auto-Encoder (VAE) [41] and its variant β-VAE [42] were representative disentanglement methods to learn latent representations using a generative model. Based on VAE, some research has been proposed for bivariate disentanglement [22]. For example, both the Cross-domain Disentanglement Network (CdDN) [43] and the Interaction Information Auto-Encoder (IIAE) proposed a cross-domain disentanglement to learn domain-specific and domain-shared representations for the image-to-image translation task. Unlike VAE, some works, such as Information Maximizing Generative Adversarial Networks (InfoGAN) [44] and Disentangled Graph Contrastive Learning (DGCL) [40], adopted Generative Adversarial Network (GAN) [45] and contrastive learning to learn disentangled representations by maximizing mutual information [46] between latent variables and inputs. As for spatial-temporal forecasting, Spatial-Temporal Normalization (ST-Norm) [47] proposed to disentangle observed data into a high-frequency component and a local component.

Although bivariate disentanglement methods have achieved promising success, they are difficult to generalize into multivariate scenarios. With the number of variates increasing, it raises an issue of how to determine the effect of one disentangled representation on the others. A feasible solution is to estimate the mutual information between different variables, but the optimization of multivariate disentanglement with mutual information remains a challenging problem due to the intractable posterior distribution of variables. Due to this, we propose a joint training model and derive a lower bound estimator to optimize multivariate disentanglement with mutual information evaluation.

III. Preliminaries

In this paper, we focus on traffic flow forecasting and briefly revisit the definition and notation of traffic flow forecasting.

Definition 1 (**Spatial Region**). *There are plenty of definitions to model the regions in a city. In this study, we follow the conventional grid definition [20] that partitions a city into $H \times W$ grid maps with the same size based on longitude and latitude, such that each grid represents a spatial region $r_{h,w}$ ($h \in [1, \cdots, H]$, $w \in [1, \cdots, W]$). The grid map takes into account the traffic conditions in regions and helps to design regions' traffic scheduling and management. For example, bike-sharing companies can use regions' traffic volumes to decide how many bikes should be placed in these regions.*

Definition 2 (**Inflow/Outflow**). *After the grid-based partition, we represent the distributions of traffic volume between regions at the i-th time interval as a tensor $X_i \in \mathbb{R}^{2\times H\times W}$ where $(X_i)_{0,h,w} = x_i^{0,h,w}$ and $(X_i)_{1,h,w} = x_i^{1,h,w}$ denote the outflow and inflow volumes for a region (h, w), respectively. Formally, the outflow and inflow volumes are defined respectively as,*

$$x_i^{0,h,w} = \sum_{M_{r_k}\in\mathbb{P}} \left|\{i > 1 | u_{i-1} \in (h,w) \cap u_i \notin (h,w)\}\right|, \quad (1)$$

$$x_i^{1,h,w} = \sum_{M_{r_k}\in\mathbb{P}} \left|\{i > 1 | u_{i-1} \notin (h,w) \cap u_i \in (h,w)\}\right|, \quad (2)$$

where $|\cdot|$ denotes the cardinality of a set. $\mathbb{P}$ represents the collection of trajectories at the i^{th} time interval. $M_r : u_1 \to u_2 \to \cdots \to u_{|M_r|}$ is a trajectory in $\mathbb{P}$, and $u_k \in (h, w)$ means that a spatial point u_k lies within region (h, w), and vice versa.

Definition 3 (**Closeness/Period/Trend**). *To study the multi-periodicity of traffic, a temporal sequence of traffic flow can be intercepted into three sub-series with different resolutions, i.e., C (closeness), P (period), and T (trend). In this paper, we choose the hourly, daily, and weekly resolutions to represent the closeness, period, and trend sub-series because the traffic flow usually changes rapidly. Suppose that the sampling frequency is f times per day, and the lengths of C, P, and T are L_c, L_p, and L_t, respectively. Closeness, period, and trend sub-series for the i-th time interval can be defined as follows:*

$$C_i = [X_{i-L_c+1}, X_{i-L_c+2}, \cdots, X_i], \quad (3)$$

$$P_i = \left[X_{i-L_p\times f}, X_{i-(L_p-1)\times f}, \cdots, X_{i-1\times f},\right] \quad (4)$$

$$T_i = \left[X_{i-L_t\times f\times 7}, X_{i-(L_t-1)\times f\times 7}, \cdots, X_{i-1\times f\times 7}\right]. \quad (5)$$

Notably, the multi-periodicity (i.e., closeness, period, and trend) can also be defined as other resolutions depending on different forecasting requirements, such as {minutely, hourly, daily} for short-term forecasting and {daily, monthly, yearly} for long-term forecasting.

Definition 4 (**Traffic Flow Prediction**). *Given the historical observations* $\{X_i | i = 0, \cdots, n-1\}$*, the target of one-step traffic flow prediction is to find a model* $\mathcal{F}$ *that uses a multi-periodic subset of observations to predict the inflow and outflow volumes of regions at the next timestamp,*

$$Y_n = \mathcal{F}\left(C_{n-1}, P_{n-1}, T_{n-1}\right), \tag{6}$$

and the multi-step traffic flow prediction aims to use several multi-periodic subsets of observations to predict the inflow and outflow volumes of regions at the next l*-timestep.*

$$\sum_{j=n}^{n+l-1} Y_j = \mathcal{F}\left(\sum_{j=n-l}^{n-1} C_{n-j}, P_{n-j}, T_{n-j}\right) \tag{7}$$

IV. Methodology

Our MUSE-Net first proposes Disentanglement, Semantic-Pushing, and Semantic-Pulling modules to model the temporal multi-periodicity. Then, an existing ResPlus network [20] is adopted to capture spatial dependency. After that, we introduce an optimization and joint training procedure to solve the disentanglement problem for traffic flow forecasting.

A. Disentanglement

To address the limitations of entanglement learning, we aim to disentangle the flow of closeness, period, and trend sub-series into corresponding exclusive representations along with an interactive representation. Each exclusive representation is to capture the private property of the corresponding time sub-series, which can be useful to model the level shift and point shift of a time series. In addition, interactive representation with common patterns of traffic flow is to reduce the semantic gaps among multiple time series, which can be essential to alleviate the problem of interaction shift. That is, the exclusive representation can characterize the traffic dynamics during peak periods, while the interactive representation can describe the traffic steadiness during non-peak periods.

We assume that a set of ternary time sub-series is generated by some random process, i.e., $(c, p, t) \sim q_D(c, p, t)$, where each element of a triplet $c \in C$, $p \in P$ and $t \in T$ is respectively extracted from closeness, period and trend sub-series, and $q_D(\cdot)$ is an unknown true joint distribution. This temporal triplet can be factorized into four parts, including exclusive representations $z^c \in Z^C$, $z^p \in Z^P$ and $z^t \in Z^T$, and interactive representation $z^s \in Z^S$, which can be rewritten as a marginal likelihood maximization problem [41],

$$\begin{aligned} &\max \mathcal{L}_{dis} = \max q_\theta(c, p, t) \\ =&\max \int dz^c dz^p dz^t dz^s q_{\theta_c}\left(c|z^c, z^s\right) q_{\theta_p}\left(p|z^p, z^s\right) \\ &\quad q_{\theta_t}\left(t|z^t, z^s\right) q\left(z^c\right) q\left(z^p\right) q\left(z^t\right) q\left(z^s\right), \end{aligned} \tag{8}$$

where $q_\theta(c, p, t)$ is a generative distribution to approximate the unknown true distribution $q_D(c, p, t)$, and θ is a parameter of model.

B. Semantic-Pushing

Although we disentangle multiple time sub-series into exclusive and interactive representations, we cannot ensure that none of the flow patterns is shared across any disentangled representations. To address this, we propose to push the interactive representation away from arbitrary exclusive representations, such that each representation is semantically independent. To achieve this, we minimize the mutual information between each exclusive representation and interactive representation, which is equivalent to a maximization problem as follows,

$$\max \mathcal{L}_{push} = \max\left(\mathcal{L}^c_{push} + \mathcal{L}^p_{push} + \mathcal{L}^t_{push}\right), \tag{9}$$

where $\mathcal{L}^c_{push} = -I\left(Z^C; Z^S\right)$, $\mathcal{L}^p_{push} = -I\left(Z^P; Z^S\right)$ and $\mathcal{L}^t_{push} = -I\left(Z^T; Z^S\right)$ denotes the mutual information about closeness, period and trend sub-series, respectively. To get a better insight into how the mutual information work for disentangled representations, we take $\mathcal{L}^c_{push}$ as an example and rewrite mutual information between Z^C and Z^S with the help of interaction information [48]:

$$\begin{aligned} \mathcal{L}^c_{push} =& -I\left(Z^C; Z^S\right) \\ =& -I\left(C; Z^C\right) + I\left(Z^C; C|Z^S\right) - I\left(Z^C; Z^S|C\right). \end{aligned} \tag{10}$$

Due to the fact that Z^S learns from C, we have $q\left(z^c|c\right) = q\left(z^c|c, z^s\right)$. Thus, the last term in the above equation disappears, i.e., $I\left(Z^C; Z^S|C\right) = H\left(Z^C|C\right) - H\left(Z^C|C, Z^S\right) = 0$, which yields

$$\mathcal{L}^c_{push} = -I\left(C; Z^C\right) - I\left(C; Z^S\right) + I\left(C; Z^C, Z^S\right). \tag{11}$$

In Eq. (11), the first two terms are against the last term in terms of the total amount of information in Z^C and Z^S, making the Z^C and Z^S to capture the mutually exclusive information of closeness sub-series C. In addition, the mutual information $\mathcal{L}^p_{push}$ and $\mathcal{L}^t_{push}$ can be obtained just like $\mathcal{L}^c_{push}$ and will not be described in detail here due to space limitations.

C. Semantic-Pulling

To enable the learned interactive representation to fully capture the common pattern of traffic flow shared across multiple time sub-series, we propose to pull the interactive representation towards the original closeness, period, and trend sub-series. To this end, we quantify the amount of shared information among interactive representation and multiple time sub-series by maximizing interaction information [48],

$$\max \mathcal{L}_{pull} = \max I\left(C; P; T; Z^S\right) \tag{12}$$

Since Eq. (12) holds symmetry, we can rewrite the interaction information $I\left(C; P; T; Z^S\right)$ as the following objectives about closeness, period and trend sub-series, respectively [49],

$$\begin{aligned} \mathcal{L}^c_{pull} =& I\left(C; Z^S\right) - I\left(C; Z^S|P\right) - I\left(C; Z^S|T\right) + \\ & I\left(C; Z^S|P, T\right). \end{aligned} \tag{13}$$

$$\begin{aligned} \mathcal{L}^p_{pull} =& I\left(P; Z^S\right) - I\left(P; Z^S|C\right) - I\left(P; Z^S|T\right) + \\ & I\left(P; Z^S|C, T\right) \end{aligned} \tag{14}$$

$$\begin{aligned}\mathcal{L}_{pull}^{t} =& I\left(T; Z^S\right) - I\left(T; Z^S|C\right) - I\left(T; Z^S|P\right) + \\ & I\left(T; Z^S|C, P\right).\end{aligned} \tag{15}$$

It can be seen that Eq. (13) consists of four terms. The first term encourages Z^S to learn information from C. The second and third terms are to discard some information in Z^S separately related to P and T. The last term is to recover some information that is repeatedly discarded by the second and third items. To jointly consider the closeness, period, and trend sub-series, the objective of semantic-pulling can be reformulated as follows,

$$\begin{aligned}\max \mathcal{L}_{pull} =& \max 3 \cdot I\left(C; P; T; Z^S\right) \\ =& \max\left(\mathcal{L}_{pull}^{c} + \mathcal{L}_{pull}^{p} + \mathcal{L}_{pull}^{t}\right)\end{aligned} \tag{16}$$

D. Optimization

Our goal is to predict future traffic flow based on disentangled exclusive and interactive representations. To achieve this, we train the MUSE-Net to predict future traffic flow via a regression loss $\mathcal{L}_{reg}$ that minimizes the difference between predictive values Y_n and true values X_n, along with disentangling under the regularizations of semantic-pushing and -pulling. Combining Eqs. (8), (9), (16) and regression loss, we can derive the following overall objective for the multivariate disentanglement problem,

$$\max_{q_\theta, r_\phi} \mathcal{L}_{Dis} + \lambda\left(\mathcal{L}_{Push} + \mathcal{L}_{Pull}\right) - \mathcal{L}_{Reg} \tag{17}$$

where λ is a balanced parameter to trade off the amount of information captured by interactive representation with that from exclusive representation. It can be seen that the proposed objective function is significantly different from the existing disentangle-based methods. On the one hand, the proposed disentanglement considers the multivariate scenario, making the disentanglement more practicable. On the other hand, the proposed method adopts mutual information to quantify the information among disentangled representations, encouraging the discrimination of disentanglement. However, it is intractable to directly optimize Eq. (17) since the mutual information-based regularization terms (i.e., semantic-pushing and -pulling) bring several intractable integrals to the multivariate disentanglement. Therefore, we theoretically derive a lower bound estimator to straightforwardly differentiate and optimize the multivariate disentanglement problem as follows.

Optimizing $\mathcal{L}_{dis}$. It is intractable to solve $\mathcal{L}_{dis}$ (i.e., Eq. (8)) since the interactive representation z^s and the true parameter θ^* are unknown. Inspired by [50], Eq. (8) can be rewritten as a lower bound on the marginal likelihood and be optimized via variational inference:

$$\begin{aligned}&\mathcal{L}_{dis} = \log q_\theta\left(c, p, t\right) \\ \geq& \mathbb{E}_{r_\phi\left(z^c, z^p, z^t, z^s|c,p,t\right)}\left[\log \frac{q_\theta\left(c, p, t, z^c, z^p, z^t, z^s\right)}{r_\phi\left(z^c, z^p, z^t, z^s|c, p, t\right)}\right],\end{aligned} \tag{18}$$

where $r_\phi\left(z^c, z^p, z^t, z^s|c, p, t\right)$ is an approximated posterior for true posterior distribution $q_\theta\left(z^c, z^p, z^t, z^s|c, p, t\right)$, and can be calculated as follows:

$$\begin{aligned}&r_\phi\left(z^c, z^p, z^t, z^s|c, p, t\right) \\ =& r_\phi\left(z^c|c\right) r_\phi\left(z^p|p\right) r_\phi\left(z^t|t\right) r_\phi\left(z^s|c, p, t\right).\end{aligned} \tag{19}$$

Thus, we reformulate Eq. (18) into the following inequation:

$$\begin{aligned}\mathcal{L}_{dis} \geq & \sum_{i\in\{c,p,t\}} \mathbb{E}_{r_\phi(z^i|i) r_\phi(z^s|c,p,t)}\left[\log q_\theta\left(i|z^i, z^s\right)\right] \\ & - \sum_{i\in\{c,p,t\}} D_{KL}\left[r_\phi\left(z^i|i\right) \| q_\theta\left(z^i\right)\right] \\ & - D_{KL}\left[r_\phi\left(z^s|c,p,t\right) \| q_\theta\left(z^s\right)\right].\end{aligned} \tag{20}$$

Optimizing $\mathcal{L}_{push}$. $\mathcal{L}_{push}$ (i.e., Eq. (9)) is intractable due to the unknown distribution $q_D\left(c\right)$, $q_D\left(p\right)$ and $q_D\left(t\right)$; therefore, we apply the Variational Information Bottleneck (VIB) [51] to simplify and optimize Eq. (9). Taking the objective of semantic-pushing about closeness as an example (i.e., Eq. (11)), the first term $-I\left(C; Z^C\right)$ and the second term $-I\left(C; Z^S\right)$ can be approximated by using $-\mathbb{E}_{q_D(c)}\left[D_{KL}\left[r_\phi\left(z^c|c\right) \| |q_\theta\left(z^c\right)\right]\right]$ and $-\mathbb{E}_{q_D(c)}\left[D_{KL}\left[r_\phi\left(z^s|c\right) \| |q_\theta\left(z^s\right)\right]\right]$, respectively, as their lower bound, where $q_\theta\left(z^c\right)$ and $q_\theta\left(z^s\right)$ can be defined as the standard Gaussian distribution. In addition, we can maximize the lower bound of the last term $I\left(C; Z^C, Z^S\right)$ by using a generative distribution $q_\theta\left(c|z^c, z^s\right)$ as follows,

$$\begin{aligned}&I\left(C; Z^C, Z^S\right) \\ =& E_{r_\phi(z^c,z^s|c)q_D(c)}\left[\log \frac{r_\phi\left(c|z^c, z^s\right)}{q_D(c)}\right] \\ =& H\left(C\right) + \mathbb{E}_{r_\phi(z^c,z^s|c)q_D(c)}\left[q_\theta\left(c|z^c, z^s\right)\right] + \\ & \mathbb{E}_{r_{\phi(z^c,z^s)}}\left[D_{KL}\left[r_\phi\left(c|z^c, z^s\right) \| q_\theta\left(c|z^c, z^s\right)\right]\right] \\ \geq& H\left(C\right) + \mathbb{E}_{r_\phi(z^c,z^s|c)q_D(c)}\left[q_\theta\left(c|z^c, z^s\right)\right].\end{aligned} \tag{21}$$

Thus, the objective of semantic-pushing about closeness sub-series can be derived as follows,

$$\begin{aligned}\mathcal{L}_{push}^{c} \geq & -\mathbb{E}_{q_D(c)}\left[D_{KL}\left[r_\phi\left(z^c|c\right) \| |q_\theta\left(z^c\right)\right]\right] \\ & -\mathbb{E}_{q_D(c)}\left[D_{KL}\left[r_\phi\left(z^s|c\right) \| |q_\theta\left(z^s\right)\right]\right] \\ & +\mathbb{E}_{r_\phi(z^c,z^s|c)q_D(c)}\left[q_\theta\left(c|z^c, z^s\right)\right].\end{aligned} \tag{22}$$

Optimizing $\mathcal{L}_{pull}$. Similar to the optimization of $\mathcal{L}_{push}$, we apply the VIB [51] technique to optimize the intractable $\mathcal{L}_{pull}$ (that is, Eq. (16)). Taking the objective of semantic-pulling about closeness as an example (i.e., Eq. (13)), the lower bound of last term $I\left(C; Z^S|P, T\right)$ can be maximized via a variational distribution $d_\omega^{p,t}\left(z^s|p, t\right)$ for paired sub-series P and T as follows,

$$\begin{aligned}&I\left(C; Z^S|P, T\right) \\ =& \mathbb{E}_{q_D(c,p,t)r_\phi(z^s|c,p,t)}\left[\log \frac{r_\phi\left(z^s|c, p, t\right)}{r_\phi\left(z^s|p, t\right)}\right] \\ =& \mathbb{E}_{q_D(c,p,t)r_\phi(z^s|c,p,t)}\left[\log \frac{r_\phi\left(z^s|c, p, t\right) d_\omega^{p,t}\left(z^s|p, t\right)}{d_\omega^{p,t}\left(z^s|p, t\right) r_\phi\left(z^s|p, t\right)}\right] \\ =& \mathbb{E}_{q_D(c,p,t)}\left[D_{KL}\left[r_\phi\left(z^s|c, p, t\right) \| d_\omega^{p,t}\left(z^s|p, t\right)\right]\right] + \\ & \mathbb{E}_{q_D(c,p,t)}\left[D_{KL}\left[d_\omega^{p,t}\left(z^s|p, t\right) \| r_\phi\left(z^s|p, t\right)\right]\right] \\ \geq& \mathbb{E}_{q_D(c,p,t)}\left[D_{KL}\left[r_\phi\left(z^s|c, p, t\right) \| d_\omega^{p,t}\left(z^s|p, t\right)\right]\right].\end{aligned} \tag{23}$$

For the second term $I\left(C; Z^S|P\right)$ and third term $I\left(C; Z^S|T\right)$ in Eq. (13), we can derive the following lower

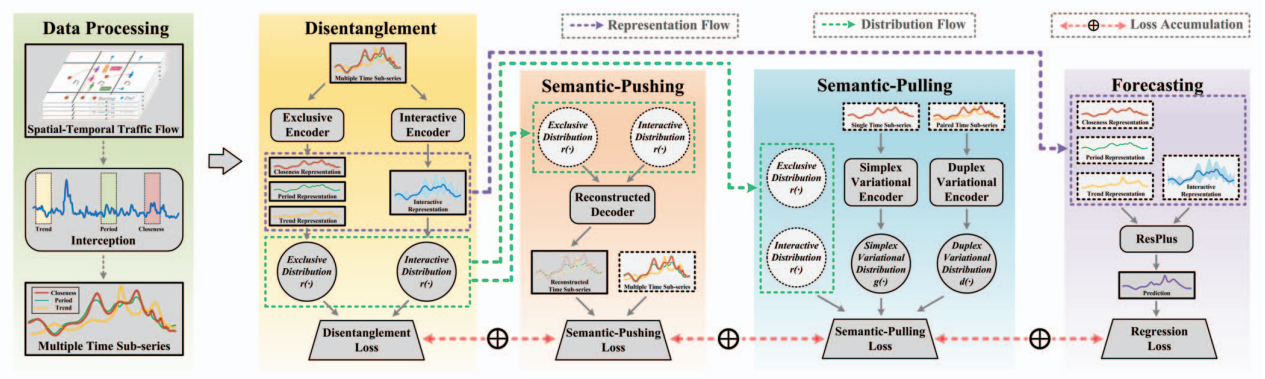


Fig. 3. The framework of our proposed MUSE-Net. Specifically, we first intercept data into ternary time sub-series, including closeness C, period P, and trend T sub-series. Then, we disentangle the ternary time sub-series into exclusive and interactive representations and distributions via exclusive and interactive encoders, respectively. After that, we take the exclusive and interactive distributions as the input of the reconstructed decoder to decode and obtain reconstructed ternary time sub-series, which can be used for semantic pushing. Meanwhile, we utilize simplex and duplex variational encoders to obtain variational distributions. Variational distributions are combined with exclusive and interactive distributions for semantic pulling. Finally, we fuse exclusive and interactive representations by an existing ResPlus network to capture spatial dependency and predict unknown traffic flow.

bound of the second term by introducing a variational distribution $g_\tau^p\left(z^s|p\right)$, and the third term is analogous:

$$\begin{aligned}
&-I\left(C;Z^S|P\right)\\
=&-\mathbb{E}_{q_D(c,p)r_\phi(z^s|c,p)}\left[\log\frac{r_\phi\left(z^s|c,p\right)}{r_\phi\left(z^s|p\right)}\right]\\
=&-\mathbb{E}_{q_D(c,p)r_\phi(z^s|c,p)}\left[\log\frac{r_\phi\left(z^s|c,p\right)d_\omega^{c,p}\left(z^s|c,p\right)g_\tau^p\left(z^s|p\right)}{d_\omega^{c,p}\left(z^s|c,p\right)g_\tau^p\left(z^s|p\right)r_\phi\left(z^s|p\right)}\right]\\
=&-\mathbb{E}_{q_D(c,p)}\left[D_{KL}\left[d_\omega^{c,p}\left(z^s|c,p\right)\|g_\tau^p\left(z^s|p\right)\right]\right]\\
&+\mathbb{E}_{q_D(c,p)}\left[D_{KL}\left[d_\omega^{c,p}\left(z^s|c,p\right)\|r_\theta\left(z^s|c,p\right)\right]\right]\\
&+\mathbb{E}_{q_D(c,p)}\left[D_{KL}\left[r_\theta\left(z^s|p\right)\|g_\tau\left(z^s|p\right)\right]\right]\\
\geq&-\mathbb{E}_{q_D(c,p)}\left[D_{KL}\left[d_\omega^{c,p}\left(z^s|c,p\right)\|g_\tau^p\left(z^s|p\right)\right]\right].
\end{aligned}\tag{24}$$

Thus, we can derive the objective of semantic-pulling about closeness sub-series as follows,

$$\begin{aligned}
\mathcal{L}_{pull}^c\geq&\mathbb{E}_{q_D(c)}\left[D_{KL}\left[r_\phi\left(z^s|c\right)\|\|q_\theta\left(z^s\right)\right]\right]-\\
&\mathbb{E}_{q_D(c,p)}\left[D_{KL}\left[d_\omega^{c,p}\left(z^s|c,p\right)\|g_\tau^p\left(z^s|p\right)\right]\right]-\\
&\mathbb{E}_{q_D(c,t)}\left[D_{KL}\left[d_\omega^{c,t}\left(z^s|c,t\right)\|g_\tau^t\left(z^s|t\right)\right]\right]+\\
&\mathbb{E}_{q_D(c,p,t)}\left[D_{KL}\left[r_\phi\left(z^s|c,p,t\right)\|d_\omega^{p,t}\left(z^s|p,t\right)\right]\right].
\end{aligned}\tag{25}$$

Overall Objective Function. After obtaining the lower bound of $\mathcal{L}_{dis}$, $\mathcal{L}_{push}$ and $\mathcal{L}_{pull}$, we can reformulate the overall objective function (i.e., Eq. (17)) by merging and canceling out terms,

$$\max_{q_\theta,r_\phi}\widehat{\mathcal{L}}_{Dis}+\widehat{\mathcal{L}}_{Push}+\widehat{\mathcal{L}}_{Pull}-\mathcal{L}_{Reg},\tag{26}$$

where

$$\begin{aligned}
\widehat{\mathcal{L}}_{Dis}=&-(1+\lambda)\cdot\mathbb{E}\left[\sum_{i\in\{c,p,t\}}D_{KL}\left[r_\phi\left(z^i|i\right)\|q_\theta\left(z^i\right)\right]\right]\\
&-\mathbb{E}_{q_D(c,p,t)}\left[D_{KL}\left[r_\phi\left(z^s|c,p,t\right)\|q_\theta\left(z^s\right)\right]\right],
\end{aligned}\tag{27}$$

$$\widehat{\mathcal{L}}_{Push}=(1+\lambda)\cdot\mathbb{E}\left[\sum_{i\in\{c,p,t\}}\log q_\theta\left(i|z^i,z^s\right)\right],\tag{28}$$

$$\begin{aligned}
\widehat{\mathcal{L}}_{Pull}=&\lambda\cdot\mathbb{E}\left[-\sum_{\substack{i,j\in\{c,p,t\}\\i\neq j}}D_{KL}\left[d_\omega^{i,j}\left(z^s|i,j\right)\|g_\tau^i\left(z^s|i\right)\right]\right.\\
&\left.+\sum_{\substack{i\in\{c,p,t\}\\i\neq j}}D_{KL}\left[r_\phi\left(z^s|c,p,t\right)\|d_\omega^{i,j}\left(z^s|i,j\right)\right]\right],
\end{aligned}\tag{29}$$

$$\mathcal{L}_{Reg}=\|X_n-Y_n\|_2^2.\tag{30}$$

E. Joint Training

The overall objective function Eq. (26) consists of four components. Specifically, Eqs. (27), (28), (29), and (30) denote disentanglement, semantic-pushing, semantic-pulling and forecasting units, respectively. To achieve this, we propose a joint training framework, as shown in Fig. 3. The details of the joint training framework are as follows.

In Eq. (27), we propose an **exclusive encoder** to learn exclusive information. The exclusive encoder first takes a time sub-series as input (e.g., C), then utilizes a convolutional layer to encode the exclusive representation (e.g., Z^C) of time sub-series, and a fully connected layer to extract the distribution of representation (e.g., $r_\phi\left(z^c|c\right)$). Meanwhile, we propose an **interactive encoder** to learn interactive information. The interactive encoder takes the convolutional features of ternary time sub-series, including C, P, and T, as inputs and consists of two components: a convolutional layer for learning interactive representation (i.e., Z^S), and a fully connected layer for learning corresponding distribution (i.e., $r_\phi\left(z^s|c,p,t\right)$).

In Eq. (28), we take the generative distributions (e.g., $q_\theta\left(c|z^c,z^s\right)$) as a **reconstructed decoder**. The reconstructed decoder aims to reconstruct a time sub-series (e.g., $\widehat{C}$) based on corresponding exclusive (e.g., Z^C) and interactive (i.e., Z^S) representations by using a fully connected layer.

In Eq. (29), we propose a **simplex variational encoder** to approximate the variational distribution for single time sub-series (e.g., $g_\tau^p\left(z^s|p\right)$), and a **duplex variational encoder** to

TABLE I
THE COMPARISON OF TIME AND SPACE COMPLEXITY OF DIFFERENT METHODS.

Method	Class	Type	Complexity
DeepSTN+ [20]	CNN	Time	$\mathcal{O}(LdM + d^2M + dM^2)$
		Space	$\mathcal{O}(Ld + d^2 + dM^2))$
DMSTGCN [52]	GCN	Time	$\mathcal{O}(Ld^2M + LdE)$
		Space	$\mathcal{O}(LdM + d^3 + M^2)$
GMAN [38]	Attention	Time	$\mathcal{O}(Ld^2M + LdM^2)$
		Space	$\mathcal{O}(LdM + L^2M + LM^2 + d^2)$
MUSE-Net (Ours)	CNN	Time	$\mathcal{O}(LdM + d^2M + dM^2)$
		Space	$\mathcal{O}(Ld + d^2 + dM^2))$

approximate the variational distribution for paired time sub-series (e.g., $d_{\omega}^{p,t}\left(z^s|p,t\right)$). The simplex variational encoder takes the convolutional feature of a time sub-series (e.g., P) as input and extracts the variational distribution by the combination of a convolutional layer and a fully connected layer. The duplex variational encoder is similar to the simplex one but takes paired time sub-series (e.g., P and T) as inputs.

In Eq. (30), we aim to fit the prediction Y_n into real future traffic flow X_n based on the learned exclusive and interactive representations. Therefore, we adopt a **ResPlus** network proposed by DeepSTN+ [20] that is designed to model spatial dependency, fuse exclusive and interactive representations, and generate the prediction of future traffic flow Y_n.

Following DeepSTN+ [20], we set the lengths of the closeness, period, and trend subseries (that is, L_c, L_p and L_t) to 3, 4, and 4 steps. The dimension of learned exclusive and interactive representations is set to $d = 64$. The distribution of representation is denoted by mean and standard deviation. Empirically, we sample the mean and standard deviation with dimension $k/4$ from the exclusive representations while sampling the mean and standard deviation with the dimension k from the interactive representation, where $k = 128$.

F. Complexity Analysis

Table I tabulates the comparison of time and space complexity of the proposed MUSE-Net with representative CNN-based, GCN-based, and Attention-based baselines, including DeepSTN+ [20], DMSTGCN [52], and GMAN [38], where $L = L_c + L_p + L_t$, d, $M = H \times W$, and E denote the length of multi-periodic series, the representation dimension, the grid size, and the number of edges in a graph, respectively. Since the proposed MUSE-Net is mainly dependent on convolution, the time complexity of MUSE-Net is $\mathcal{O}(LdM + d^2M)$. As shown in Table I, the MUSE-Net can be faster than GMAN because $L, d \ll M$. Moreover, if the graph is dense, i.e., $E \to M^2$, the time complexity of DMSTGCN [52] will be higher than the proposed method. In terms of space complexity, although the proposed MUSE-Net requires slightly more memory than DMSTGCN and GMAN, the space complexity of MUSE-Net has the same magnitude as the baselines becasue these methods all have M^2 space complexity. Considering the superior performance of the MUSE-Net achieved (please see Table II in the Experiments Section), the space complexity of the proposed method is acceptable.

V. EXPERIMENTS

In this section, we evaluate our proposed MUSE-Net on three public benchmark datasets in comparison with the state-of-the-art traffic flow forecasting methods, which are summarized to answer the following research questions:

- RQ1: Does our proposed MUSE-Net outperform baselines in traffic flow forecasting?
- RQ2: Does the design of different components contribute to the performance of the model?
- RQ3: Are disentangled exclusive and interactive representations independent of each other?
- RQ4: Can exclusive and interactive representations provide sufficient information for forecasting?
- RQ5: Can exclusive and interactive representations interpret specific traffic flow patterns?
- RQ6: How do the hyper-parameters of MUSE-Net affect the performance of the prediction task?

A. Datasets

We evaluate the proposed method on three real-world benchmark datasets, as detailed follows,

- NYC-Bike [8]: The NYC-Bike dataset consists of bike trajectories in New York from 07/01/2016 to 08/29/2016. Following [53], we first divide the entire city as grid maps of 10×20. The size of each grip is about $1km \times 1km$. Then, we select data from the first 40 days (i.e., from 07/01/2016 to 08/09/2016) as the training set and the data from the last 20 days as the testing set.
- NYC-Taxi [8]: The NYC-Taxi dataset consists of taxicab trajectories in New York from 01/01/2015 to 03/01/2015. Following [53], we first divide the entire city as grid maps of 10×20. The size of each grip is about $1km \times 1km$. Then we select data from the first 40 days (that is, from 01/01/2015 to 02/10/2015) as the training set and data from the last 20 days as the testing set.
- TaxiBJ [15]: The TaxiBJ dataset includes taxicab GPS trajectories collected from 01/01/2013 to 10/30/2013. Following [16], we first divide the entire city as grid maps of 32×32. The size of each grip is about $0.6km \times 0.6km$. Then, we select data from the last 20 days (i.e., from 10/11/2013 to 10/30/2013) as the testing set and the remaining data as the training set.

In our experiment, the length of each time interval is set to 30 minutes. We use tanh as our final activation function of which output ranges between -1 and 1. Thus, we scale the data into the range $[-1, 1]$ via the Min-Max normalization during training and re-scale the predicted value back to the normal values for comparison with the ground-truth during evaluation. Moreover, we select 90% of the training data to fit the models and the remaining 10% for validation.

TABLE II
ONE-STEP FORECASTING COMPARISON OF ALL METHODS IN THE NYC-BIKE, NYC-TAXI, AND TAXIBJ DATASETS, RESPECTIVELY.

Method	NYC-Bike						NYC-Taxi						TaxiBJ					
	Outflow			Inflow			Outflow			Inflow			Outflow			Inflow		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
RNN [7]	12.79	4.18	82.92%	13.27	4.23	84.97%	32.81	13.53	32.83%	35.24	13.32	41.52%	33.27	18.26	22.35%	37.64	18.77	23.12%
Seq2Seq [54]	11.26	4.02	75.45%	9.99	3.56	64.92%	30.26	9.45	33.75%	31.14	11.09	42.49%	27.92	16.26	20.48%	27.71	16.48	20.86%
ASTGCN [18]	6.11	1.89	40.49%	5.35	1.75	38.17%	28.96	9.31	30.79%	25.01	8.76	30.85%	21.99	13.43	20.32%	22.28	13.55	20.50%
CONVGCN [55]	3.80	1.59	25.28%	3.72	1.56	25.56%	21.23	12.28	27.45%	21.62	12.59	31.69%	18.77	11.16	16.50%	18.91	11.27	16.68%
GMAN [38]	3.63	1.34	24.45%	3.43	1.27	23.24%	22.36	7.76	23.39%	24.11	8.64	25.29%	21.39	12.93	18.63%	21.47	12.98	18.72%
STGNN [25]	6.49	1.99	40.65%	6.50	1.96	39.51%	25.52	8.65	32.93%	22.86	7.92	27.80%	21.72	13.36	19.69%	21.92	13.75	20.09%
DMSTGCN [52]	3.68	1.45	25.42%	3.82	1.86	28.98%	18.67	9.80	29.43%	18.24	7.84	23.66%	20.74	12.12	17.03%	21.09	12.25	17.57%
ST-Norm [47]	3.75	1.61	25.12%	3.59	1.55	23.94%	19.51	8.94	29.10%	17.45	8.16	27.60%	19.20	11.44	16.74%	19.90	11.95	17.78%
STGSP [16]	3.86	1.26	25.46%	3.80	1.25	25.87%	21.74	6.99	23.04%	21.60	7.38	23.83%	21.53	12.63	17.90%	21.64	12.71	18.06%
DeepSTN+ [20]	3.68	1.35	25.07%	3.48	1.33	24.54%	18.25	7.37	27.07%	18.33	7.94	33.13%	18.30	10.81	15.77%	18.38	10.87	15.88%
ST-SSL [17]	4.56	1.26	24.11%	4.07	1.22	23.46%	23.79	7.12	21.80%	24.57	7.05	21.25%	20.45	11.41	15.89%	20.47	11.42	15.92%
MUSE-Net (Ours)	**2.89**	**1.11**	**21.28%**	**2.73**	**1.06**	**20.70%**	**15.16**	**5.40**	**19.94%**	**14.05**	**5.42**	**19.47%**	**17.16**	**10.28**	**14.60%**	**17.26**	**10.35**	**14.72%**
Improvement	20%	12%	12%	20%	13%	12%	17%	23%	9%	19%	23%	8%	6%	5%	7%	6%	5%	7%

B. Baselines & Implementation

We implement the proposed model with the Keras framework and train our model using the Adam optimizer with a learning rate of 0.0002, batch size of 8, and maximal epoch of 350. For the objective function, the balance coefficient λ is set as 1 to trade off the information learning. Then, We compare MUSE-Net with the following 11 baselines, which can be grouped into five classes: RNN-based models (RNN, Seq2Seq), CNN-based models (CONVGCN, DeepSTN+), GNN-based models (ASTGCN, DMSTGCN, STGNN, ST-SSL), Attention-based models (GMAN, STGSP) and Disentangle-based model (ST-Norm).

- RNN [7]: It leverages a recurrent neural network to capture temporal effects for forecasting time series data.
- Seq2Seq [54]: It is an encoder-decoder framework using a gated recurrent neural network to predict traffic flow.
- ASTGCN [18]: It includes a spatial-temporal component where a graph convolutional network and a convolution network are to learn spatial and temporal information, respectively, and an attention component which is to capture spatial-temporal correlations.
- CONVGCN [55]: It combines a graph convolution with a 3D convolution to capture short-term and long-term spatial dependency for traffic flow forecasting.
- GMAN [38]: It designs an encoder-decoder framework with a transform attention mechanism that converts historical traffic flow into future traffic flow.
- STGNN [25]: It combines a position-wise graph neural network with a recurrent-based transformer layer to jointly capture the spatial and temporal dependency for traffic flow forecasting.
- DMSTGCN [52]: It explores the time-specific spatial dependency of traffic flow via a dynamic graph convolutional network with dilated convolution layer.
- ST-Norm [47]: It proposes temporal and spatial normalization that separately refines the high-frequency and local components to model the patterns of traffic flow.
- STGSP [16]: It is a transformer-based method that can model the dynamic correlation of multi-periodic patterns by using the multi-head attention mechanism.
- DeepSTN+ [20]: It jointly learns the temporal and spatial dependencies to predict future traffic flow via a convolutional neural network.
- ST-SSL [17]: It proposes a self-supervised learning paradigm to model spatial and temporal heterogeneities for traffic flow forecasting.

For the ASTGCN, CONVGCN, GMAN, DMSTGCN, and ST-Norm methods, we carefully modify their output channel of prediction layer from 1 to 2 to jointly predict inflow and outflow. For the STGSP method, we do not utilize external information (such as date and weather) to predict traffic flow for a fair comparison. For the rest of the baselines, we adopt their original networks and follow the best parameters reported in papers to perform the experiments.

All experiments are conducted on a Linux server (CPU: Intel(R) Xeon(R) Gold 6138 CPU @ 2.00GHz, GPU: NVIDIA Tesla V100 GPU 32GB). Additionally, we employ Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) as metrics to evaluate the prediction performance of different methods [56]. For the MAE, RMSE, and MAPE metrics, smaller values indicate the better prediction performance.

C. Performance Comparison (RQ 1)

To comprehensively evaluate the performance of our proposed MUSE-Net, we first conduct comparative experiments in both one-step forecasting and multi-step forecasting settings. Then, we further evaluate the prediction performance of the MUSE-Net during peak vs. non-peak and weekday vs. weekend periods to understand the difference in prediction for diverse traffic conditions. Finally, we visualize the forecasting results of our MUSE-Net.

One-step forecasting is a fundamental task in traffic flow forecasting. In our experimental setting, we predict the traffic flow of a time series at the next time step based on the historical multi-periodic data. Table II tabulates the performance of our MUSE-Net compared to eleven baselines in the NYC-Bike, NYC-Taxi, and TaxiBJ datasets, respectively. The percentage of improvement in Table II achieved by our MUSE-Net is defined as $\frac{Best\ baseline\ result\ -\ Ours\ result}{Best\ Baseline\ result} \times 100\%$.

TABLE III
MULTI-STEP FORECASTING COMPARISON OF FOUR METHODS IN THE NYC-BIKE, NYC-TAXI, AND TAXIBJ DATASETS, RESPECTIVELY.

Dataset	Method	Horizon 1						Horizon 2						Horizon 3					
		Outflow			Inflow			Outflow			Inflow			Outflow			Inflow		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
NYC-Bike	ST-GSP	4.06	1.32	27.70%	4.01	1.31	27.18%	3.63	1.22	25.45%	3.71	1.24	25.67%	4.13	1.34	27.95%	3.99	1.31	27.46%
	DeepSTN+	1.29	0.52	6.82%	1.33	0.54	7.15%	1.36	0.54	7.68%	1.35	0.55	7.59%	3.21	1.12	22.34%	2.96	1.08	21.44%
	ST-SSL	3.15	0.46	6.28%	2.70	0.50	7.57%	3.18	0.49	7.04%	2.77	0.54	8.46%	4.30	1.20	22.66%	3.84	1.17	22.51%
	MUSE-Net	**1.08**	**0.37**	**5.11%**	**1.09**	**0.36**	**4.93%**	**1.19**	**0.40**	**5.55%**	**1.14**	**0.39**	**5.40%**	**3.18**	**1.08**	**21.06%**	**2.92**	**1.03**	**20.63%**
	Improvement	16.28%	19.57%	18.63%	18.05%	28.00%	31.05%	12.50%	18.37%	21.16%	15.56%	27.78%	4.76%	0.93%	3.57%	5.73%	1.35%	4.63%	3.78%
NYC-Taxi	ST-GSP	18.59	6.26	22.84%	19.24	6.50	22.43%	19.46	6.54	22.85%	20.60	7.02	23.18%	22.69	7.40	25.07%	21.43	7.23	23.87%
	DeepSTN+	5.96	2.72	15.05%	6.84	3.41	18.05%	6.63	2.92	15.06%	7.68	3.66	17.93%	14.34	4.99	19.38%	13.16	5.23	21.40%
	ST-SSL	15.34	4.3	14.84%	17.37	4.49	15.02%	14.88	4.15	13.98%	17.71	4.65	15.14%	17.57	5.38	**17.83%**	19.49	5.62	**18.14%**
	MUSE-Net	**5.25**	**2.30**	**13.17%**	**5.91**	**2.66**	**13.69%**	**6.20**	**2.58**	**13.73%**	**6.54**	**2.94**	**14.57%**	**14.22**	**4.81**	18.80%	**13.00**	**4.92**	19.01%
	Improvement	11.91%	15.44%	11.25%	13.60%	21.99%	8.85%	6.49%	11.64%	1.79%	14.84%	19.67%	3.76%	0.84%	3.61%	-5.44%	1.22%	5.93%	-4.80%
TaxiBJ	ST-GSP	18.83	11.22	15.34%	18.95	11.30	15.49%	17.70	10.66	14.77%	17.94	10.81	15.00%	20.78	11.99	16.75%	20.87	12.08	16.94%
	DeepSTN+	4.34	3.15	6.63%	4.51	3.28	6.83%	4.63	3.31	6.95%	4.55	3.25	6.95%	17.96	10.49	15.22%	18.04	10.55	15.34%
	ST-SSL	6.31	3.04	4.79%	6.60	3.08	4.82%	6.37	2.96	4.50%	6.69	3.02	4.47%	19.28	10.92	15.48%	19.39	10.95	15.57%
	MUSE-Net	**3.66**	**2.32**	**4.16%**	**3.65**	**2.28**	**4.09%**	**3.54**	**2.28**	**4.13%**	**3.52**	**2.27**	**4.16%**	**17.84**	**10.20**	**14.82%**	**17.90**	**10.26**	**14.93%**
	Improvement	15.67%	23.68%	13.15%	19.07%	25.97%	15.15%	23.54%	22.97%	8.22%	22.64%	24.83%	6.94%	0.67%	2.76%	2.63%	0.78%	2.75%	2.67%

TABLE IV
PEAK VS. NON-PEAK PERFORMANCE COMPARISON OF FOUR METHODS IN THREE DATASETS, RESPECTIVELY.

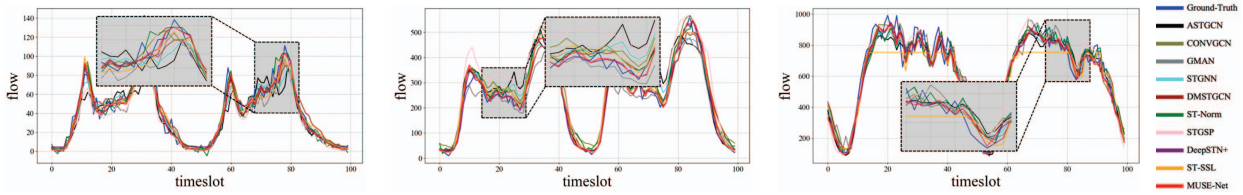
Dataset	Method	Peek				Non-peak			
		Outflow		Inflow		Outflow		Inflow	
		RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
NYC-Bike	ST-GSP	5.66	25.35%	5.26	24.81%	3.23	25.52%	3.31	26.34%
	DeepSTN+	5.30	25.82%	4.87	24.71%	2.77	24.71%	2.72	24.47%
	ST-SSL	8.20	23.92%	6.99	23.48%	2.91	24.19%	2.83	23.44%
	MUSE-Net	**4.77**	**21.14%**	**4.18**	**20.51%**	**2.42**	**21.36%**	**2.35**	**20.79%**
	Improvement	10.09%	11.63%	14.15%	12.65%	12.41%	11.71%	13.43%	11.30%
NYC-Taxi	ST-GSP	24.33	23.27%	27.55	24.55%	21.01	22.97%	19.74	23.65%
	DeepSTN+	20.10	27.25%	22.01	28.21%	16.76	27.02%	16.13	34.37%
	ST-SSL	26.27	22.56%	38.60	22.65%	23.08	21.57%	19.24	20.90%
	MUSE-Net	**17.96**	**21.54%**	**19.03**	**21.28%**	**14.49**	**19.45%**	**12.64**	**19.03%**
	Improvement	10.65%	4.52%	13.52%	6.06%	13.55%	9.81%	21.63%	8.95%
TaxiBJ	ST-GSP	22.72	16.18%	22.86	16.35%	21.20	18.40%	21.31	18.55%
	DeepSTN+	19.48	14.24%	19.62	14.36%	16.93	16.22%	16.99	16.33%
	ST-SSL	19.38	**12.69%**	19.40	**12.71%**	20.72	16.80%	20.73	16.84%
	MUSE-Net	**19.16**	13.40%	**19.25**	13.48%	**16.54**	**14.96%**	**16.63**	**15.08%**
	Improvement	1.13%	-5.59%	0.77%	-6.10%	2.29%	7.77%	2.12%	7.68%

TABLE V
WEEKDAY VS. WEEKEND PERFORMANCE COMPARISON OF FOUR METHODS IN THREE DATASETS, RESPECTIVELY.

Dataset	Method	Weekday				Weekend			
		Outflow		Inflow		Outflow		Inflow	
		RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
NYC-Bike	ST-GSP	3.88	23.62%	3.86	24.53%	3.80	30.10%	3.60	29.30%
	DeepSTN+	3.66	23.13%	3.48	22.86%	3.74	29.97%	3.47	28.88%
	ST-SSL	4.95	22.93%	4.38	22.61%	3.45	27.07%	3.21	25.63%
	MUSE-Net	**2.92**	**20.12%**	**2.72**	**19.77%**	**2.84**	**24.23%**	**2.75**	**23.10%**
	Improvement	20.22%	12.25%	21.84%	12.55%	17.68%	10.47%	14.33%	9.87%
NYC-Taxi	ST-GSP	21.28	22.24%	20.66	23.28%	22.77	24.68%	23.65	24.94%
	DeepSTN+	18.14	25.69%	18.29	32.45%	18.51	29.99%	18.42	34.56%
	ST-SSL	24.39	20.53%	25.86	20.20%	22.28	24.45%	21.24	23.41%
	MUSE-Net	**15.01**	**18.91%**	**13.78**	**18.52%**	**15.51**	**22.12%**	**14.66**	**21.46%**
	Improvement	17.25%	7.90%	24.66%	8.34%	16.21%	9.52%	20.41%	8.31%
TaxiBJ	ST-GSP	21.21	18.40%	21.32	18.56%	22.24	16.73%	22.36	16.90%
	DeepSTN+	17.94	16.22%	18.00	16.33%	19.10	14.77%	19.22	14.87%
	ST-SSL	20.13	16.19%	20.16	16.24%	21.34	15.00%	21.30	15.00%
	MUSE-Net	**16.64**	**15.00%**	**16.74**	**15.11%**	**18.31**	**13.71%**	**18.42**	**13.82%**
	Improvement	7.25%	7.35%	7.00%	6.95%	4.14%	7.12%	4.16%	7.06%

We can observe that RNN-based models, such as RNN and Seq2Seq, cannot perform well because they ignore the spatial dependency of traffic flow. For CNN-based and GNN-based models, the MUSE-Net can outperform DeepSTN+ and ST-SSL by reducing the RMSE errors by $6\% \sim 23\%$ and $16\% \sim 43\%$, respectively, in three datasets. This is because CNN-based and GNN-based models overlook the temporal sequential dependency, while the proposed method can capture the correlation between short-term and long-term time series via interactive representation. Among Attention-based models, GMAN and STGSP are capable of extracting traffic flow patterns adaptively; however, the MUSE-Net performs better than these models. Compared to STGSP, the MUSE-Net can produce $20\% \sim 35\%$ improvements in three datasets in terms of RMSE. The possible reason is that Attention-based models learn an entangled representation of traffic flow, while the MUSE-Net learns disentangled exclusive and interactive representations jointly to better characterize the traffic flow pattern. Moreover, the proposed method can achieve at least 11% improvement in RMSE compared to the disentangle-based model (i.e., ST-Norm), which verifies the rationality of MUSE-Net in modeling multi-periodicity.

In comparison to one-step forecasting, multi-step forecasting should consider not only the next step of traffic flow but also several steps later. In our experimental setting, multi-step forecasting is conducted to predict the traffic flow in 3 horizons (i.e., 1.5 hours later), and each horizon of traffic flow is characterized by corresponding historical multi-periodic traffic flow, including closeness, period, and trend data. Since several multi-periodic traffic flows are not consecutive and may exist semantic gaps, we select three multi-periodic-based methods, i.e., DeepSTN [20], ST-GSP [16], and ST-SSL [17], as baselines for comparison. Table III tabulates the multi-step forecasting results of three methods. It can be seen that our MUSE-Net obtains significant gains over the baselines in three datasets. In contrast to ST-GSP that processes multi-periodic flow sequentially, DeepSTN+, ST-SSL, and MUSE-Net separate the closeness, period, and trend flow and process each periodic flow individually, which can fully utilize specific information of each multi-periodic sub-series to obtain better performances. Moreover, DeepSTN+, ST-SSL, and MUSE-Net exploit the global information of several multi-periodic data for multi-step forecasting, such that the prediction of the first 2 horizons can be more accurate than the prediction of the



(a) NYC-Bike (b) NYC-Taxi (c) TaxiBJ

Fig. 4. Predictive results of the different methods against the ground-truth in the NYC-Bike, NYC-Taxi, and TaxiBJ datasets, respectively.

last horizon. Compared to DeepSTN+ and ST-SSL, MUSE-Net further captures exclusive and interactive traffic patterns, leading to better multi-step forecasting performance.

Since people's travel demand is different during peak and non-peak periods, as well as during weekdays and weekends, we further conduct experiments to evaluate the one-step prediction performance of the MUSE-Net compared to STGSP [16], DeepSTN+ [20], and ST-SSL [17], during the peak vs. non-peak periods and weekday vs. weekend. For the peak vs. non-peak experiment, we select the periods from 7:00 am to 9:00 am and the periods from 5:00 pm to 7:00 pm as the peak periods, and the rest of the time as the non-peak periods. For the weekday vs. weekend experiment, we select the periods from Monday to Friday as weekdays, and the rest of the time as weekends. Tables IV and Table V tabulate the peak vs. non-peak and weekday vs. weekend prediction results of four methods in the NYC-Bike, NYC-Taxi, and TaxiBJ datasets, respectively. It can be seen that the proposed MUSE-Net performs slightly worse than ST-SSL on the TaxiBJ dataset during peak periods. The possible reason is that ST-SSL may benefit from self-supervised learning with suitable augmentations and clusters. However, the proposed MUSE-Net obtains $0.77\% \sim 21.63\%$ RMSE gains and $4.14\% \sim 24.66\%$ RMSE gains over baselines for peak vs. non-peak and weekdays vs. weekends comparisons, respectively, demonstrating the robustness of MUSE-Net.

To further evaluate the effectiveness of our MUSE-Net, we illustrate the prediction of the different methods against the ground-truth in the NYC-Bike, NYC-Taxi, and TaxiBJ datasets, respectively, as shown in Fig. 4. We can observe that the proposed MUSE-Net is not only accurate in fitting the ground-truth curves during non-peak periods but also better in modeling the dynamics of ground-truth curves during peak periods. These demonstrate the proposed method's superiority over the baselines in traffic forecasting.

D. Ablation Study (RQ 2)

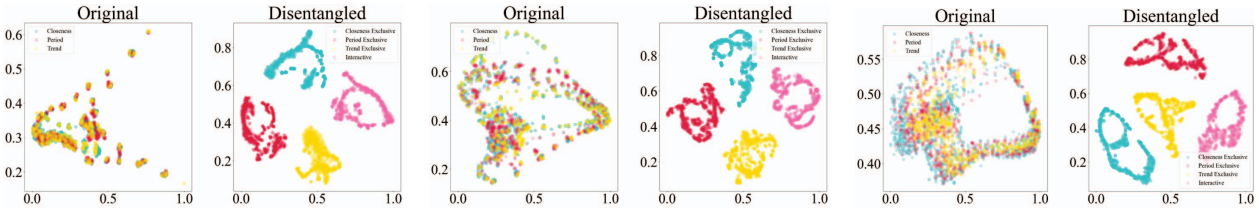
To evaluate the effectiveness and contribution of each component of the proposed MUSE-Net, we perform a comparative ablation study by implementing four variants of the MUSE-Net in three benchmark datasets:

- MUSE-Net-w/o-Spatial: drop the spatial module from our model (i.e., the model without ResPlus network)
- MUSE-Net-w/o-MultiDisentangle: use cross-variate disentanglement to replace multivariate disentanglement; that is, we learn three different interactive representations that share information across arbitrarily paired time sub-series, such as Z^{CP} sharing information across C and P, instead of one interactive representation Z^S sharing information across all time sub-series.
- MUSE-Net-w/o-SemanticPushing: drop the semantic-pushing module from our model (i.e., the overall objective function without Eq. (9))
- MUSE-Net-w/o-SemanticPulling: drop the semantic-pulling module from our model (i.e., the overall objective function without Eq. (16))

TABLE VI
THE ABLATION RESULTS OF OUR MUSE-NET IN THE NYC-BIKE, NYC-TAXI, AND TAXIBJ DATASETS, RESPECTIVELY.

Dataset	Flow	Metric	MUSE-Net -w/o- Spatial	MUSE-Net -w/o- MultiDisentangle	MUSE-Net -w/o- SemanticPushing	MUSE-Net -w/o- SemanticPulling	MUSE-Net
NYC-Bike	outflow	RMSE	3.40	3.13	2.91	2.96	**2.89**
		MAE	1.31	1.21	1.12	1.13	**1.11**
	inflow	RMSE	3.43	3.01	2.76	2.80	**2.73**
		MAE	1.32	1.19	1.08	1.08	**1.06**
NYC-Taxi	outflow	RMSE	16.28	17.22	15.57	16.05	**15.16**
		MAE	6.29	5.80	5.67	5.86	**5.40**
	inflow	RMSE	15.47	14.94	14.75	15.14	**14.05**
		MAE	6.97	5.92	5.86	6.18	**5.42**
TaxiBJ	outflow	RMSE	23.26	17.92	17.60	17.38	**17.16**
		MAE	14.11	10.82	10.63	10.45	**10.28**
	inflow	RMSE	22.98	18.02	17.66	17.45	**17.26**
		MAE	13.97	10.88	10.70	10.51	**10.35**

Table VI tabulates the ablation experimental results compared to the original MUSE-Net in the NYC-Bike, NYC-Taxi, and TaxiBJ datasets. It can be seen that our proposed MUSE-Net obviously achieves better results than its ablative variants, demonstrating the effectiveness of each part of the MUSE-Net. Moreover, we can draw the following observation. First, MUSE-Net-w/o-Spatial obtains the worst performance with $7\% \sim 35\%$ performance degradation compared to MUSE-Net, indicating the importance of spatial dependency in traffic flow forecasting. After dropping the spatial module, our MUSE-Net, which can be regarded as a temporal-only method, still achieves competitive performances over spatial-temporal methods, such as STGSP [16], STGNN [25], and GMAN [38], in three datasets (referred to Table II). This validates the effectiveness of the proposed disentanglement for traffic flow forecasting. Second, MUSE-Net-w/o-MultiDisentangle obtains the second worst performance with $4\% \sim 13\%$ performance degradation compared to MUSE-Net. By using a multivariate disentanglement module, the proposed MUSE-Net is able to directly separate the common pattern shared across different



(a) NYC-Bike (b) NYC-Taxi (c) TaxiBJ

Fig. 5. The visualized distribution of original and disentangled representations in the NYC-Bike, NYC-Taxi, and TaxiBJ datasets, respectively.

time sub-series from the private pattern existing in each time sub-series. As a result, the MUSE-Net can be more powerful in capturing the multi-periodicity of traffic flow. Third, the regularization terms, including semantic-pushing and semantic-pulling, can stably boost the traffic flow prediction due to their advantage in making the learned exclusive and interactive representations independent and informative.

E. Independence Analysis for Disentanglement (RQ 3)

The MUSE-Net attempts not only to disentangle the multi-periodic pattern into exclusive and interactive representations but also to keep each representation away from the others via semantic-pushing regularization. To validate the independence of disentanglement, we perform an experiment of 2D distribution visualization by comparing original data with disentangled representations. Specifically, we first learn the disentangled representations (i.e., three exclusive representations and one interactive representation) from the original multiple time sub-series (i.e., closeness, period, and trend sub-series) and then simultaneously project the original data and disentangled representations into the 2D distribution via t-sne [57]. In this way, we can verify independence by identifying the clusters of different representations. Fig. 5 visualizes the 2D distributions of original data and disentangled representations. It is obvious that the original data of different time sub-series are mixed up, indicating the entanglement among multiple original time sub-series. On the contrary, each disentangled representation keeps separated from the others, encouraging discrimination to capture specific patterns in the different temporal dimensions. This verifies that our proposed method can effectively disentangle multi-periodicity patterns from traffic flow and ensure the independence of each pattern.

F. Informativeness Analysis for Disentanglement (RQ 4)

Although the MUSE-Net can decouple multi-periodic patterns into exclusive and interactive representations, do these representations provide enough information for traffic flow forecasting? Here, we conduct a similarity analysis to evaluate the informativeness of disentangled representations on the TaxiBJ dataset. Specifically, we first calculate the cosine similarity between paired representations and then visualize the similarity matrices via a heatmap. The similarity value ranges from -1 to 1, which can describe how much information one representation provides for another. The higher the similarity value, the more information can be provided.

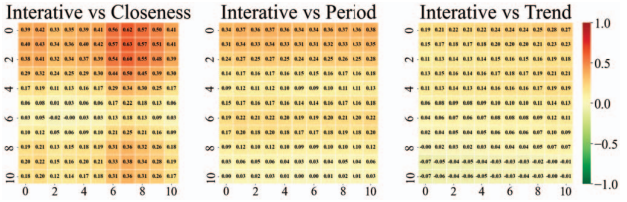


Fig. 6. The visualized similarity matrices of interactive representation with respect to original closeness, period, and trend time sub-series, respectively, in the TaxiBJ dataset.

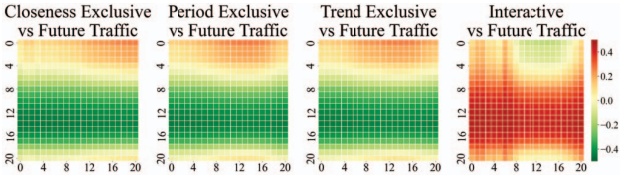


Fig. 7. The visualized similarity matrices of closeness, period, and trend exclusive representations and interactive representation with respect to future traffic flow, respectively, in the TaxiBJ dataset.

We first conduct a similarity analysis to evaluate how much information the interactive representation learns from closeness, period, and trend sub-series. To this end, we calculate and depict the similarity of interactive representation with respect to original closeness, period, and trend sub-series, as illustrated in Fig. 6. It is obvious that most points in the three heatmaps are greater than zero, indicating that interactive representation can learn enough useful information from closeness, period, and trend sub-series. This verifies the effectiveness of our proposed semantic-pulling regularization on pulling the interactive representation towards the original multiple time sub-series.

Furthermore, we conducted another similarity analysis to evaluate how exclusive and interactive representations contribute to the forecasting. Hence, we compute and visualize the similarity for exclusive and interactive representations with respect to future traffic flow. As depicted in Fig. 7, the color (i.e., similarity distribution) of the interactive representation is opposite to that of exclusive representations. This indicates that the information of interactive representation is complementary to that of exclusive representation. As a result, the incorporation of exclusive and interactive representations can provide sufficient information to accurately predict future traffic flow.

G. Interpretability of Disentangled Representations (RQ 5)

Since disentanglement provides the interpretation power of representation learning, we further evaluate and demonstrate the meanings of disentangled exclusive and interactive representations. Specifically, we calculate the similarity matrix between disentangled representations and future traffic flow. The diagonal of the matrix represents the similarity between the future traffic flow at time t and the corresponding representations. The value of similarity ranges from -1 to 1. The larger the value, the more similar the disentangled representation and the future traffic flow will be.

Fig. 8 depicts an example of traffic flow in the region (5, 4) of the TaxiBJ dataset. This traffic flow is collected from 6:30 pm on 11/10/2013 to 9:30 am on 12/10/2013, which can be broken into peak periods (i.e., 6:30 pm-11:00 pm and 7:00 am-9:30 am) and non-peak periods (i.e., 11:30 pm-6:30 am). It can be seen that the similarity values of three exclusive representations are greater than zero during peak periods while being less than zero during non-peak periods. This observation indicates that exclusive representation reveals the traffic pattern during peak periods. The possible reason is that exclusive representation characterizing the private property is powerful in modeling the unique data distribution, such as fluctuated traffic flow. In addition, we notice that the similarity values of interactive representation in non-peak periods are greater than those in peak periods, indicating that interactive representation tends to reveal the traffic pattern during non-peak periods. The possible reason is that interactive representation is designed to capture common patterns sharing a similar data distribution. Thus, interactive representation can better model the normal distribution, such as traffic flow during non-peak periods.

H. Parameter Sensitivity (RQ 6)

The proposed MUSE-Net mainly contains three parameters, i.e., the trade-off parameter λ, the sampled dimensions of mean and standard deviation k, and the representation dimension d. To evaluate these three parameters, we first set λ, k, and k as the values from candidate set $\{10^{-3} \sim 10^{3}\}$, $\{16 \sim 1024\}$, and $\{16 \sim 320\}$, respectively. Then, we repeat the experiment ten times and take the average results as the experimental results to compare the parameters' effects on our MUSE-Net. Fig. 9 shows the RMSE values of MUSE-Net versus different parameter values on the NYC-Bike dataset, where the blue curve denotes the average results, and the light blue background denotes the fluctuation range of the results. As shown in Fig. 9 (a), the prediction performance becomes unstable when λ is much greater or much lower than 1. This is because λ trades off the amount of information captured by interactive representation with that from exclusive representation. As λ increases, MUSE-Net learns limited information. As λ decreases, MUSE-Net learns redundant information. Due to this, we empirically set $\lambda = 1$ for all datasets to obtain the best performance. From Fig. 9 (b), we can observe that the MUSE-Net can achieve comparable prediction performance over a wide range of k. The possible reason is that the mean and standard deviation with small dimensions are sufficient to represent the distributions of representations and to evaluate the difference between various representations. Therefore, we empirically set $k = 128$ for all datasets to obtain the best performance. From Fig. 9 (c), it can be seen that the proposed method is not sensitive to the parameter d and different values of the representation dimension can achieve competitive performance. Hence, we choose the parameter with the best performance as the representation dimension, that is, $d = 64$.

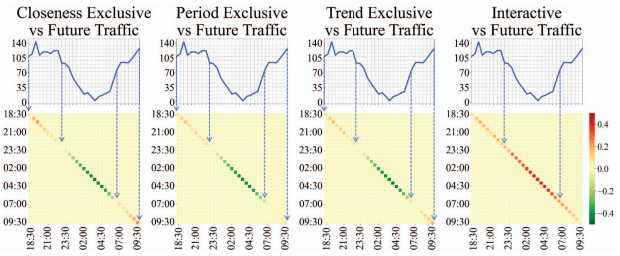
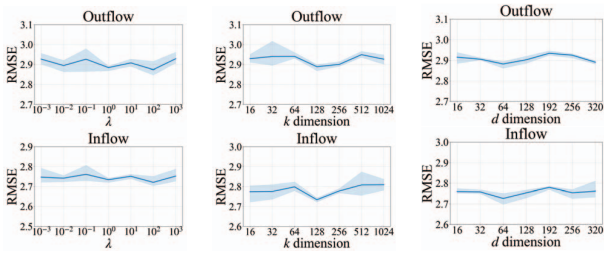


Fig. 8. An example to interpret the traffic pattern meanings of closeness, period and trend exclusive representations, and interactive representation for forecasting, respectively.



(a) Parameter λ (b) Parameter k (c) Parameter d

Fig. 9. The RMSE results of our MUSE-Net versus different values of the parameters λ, k and d in the NYC-Bike dataset.

VI. Conclusion

In this work, we investigate traffic flow forecasting by proposing a novel disentanglement framework, namely MUSE-Net, to alleviate distribution shift and interaction shift problems. In particular, our MUSE-Net can not only handle the disentangling under multiple variables but also encourage the disentangled representations to be independent and informative. Moreover, we derive a lower bound estimator to straightforwardly differentiate and optimize the disentanglement problem. Comprehensive experimental results verify the effectiveness of MUSE-Net in disentangling and demonstrate the superiority of MUSE-Net over state-of-the-art traffic flow forecasting methods.

Following mapping the data sensors to grids based on their coordinates and intercepting the data into closeness, period, and trend series, we can easily apply the proposed method to other forecasting applications, such as population-level epidemic forecasting, air-quality forecasting, and energy forecasting. Considering the power of disentanglement in multi-periodicity learning, we believe that the proposed MUSE-Net can be useful for a variety of forecasting applications.

REFERENCES

[1] J. P. Zhang, F. Y. Wang, K. F. Wang, W. H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, 2011.

[2] X. Song, Q. S. Zhang, Y. Sekimoto, and R. Shibasaki, "Prediction of human emergency behavior and their mobility following large-scale disaster," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 5–14.

[3] J. B. Zhang, Y. Zheng, D. K. Qi, R. Y. Li, and X. W. Yi, "Dnn-based prediction model for spatio-temporal data," in *Proc. ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2016, pp. 1–4.

[4] H. Lee, C. Park, S. Jin, H. Chu, J. Choo, and S. Ko, "An empirical experiment on deep learning models for predicting traffic data," in *Proc. IEEE International Conference on Data Engineering*, 2021, pp. 1817–1822.

[5] S. V. Kumar and L. Vanajakshi, "Short-term traffic flow prediction using seasonal arima model with limited input data," *European Transport Research Review*, vol. 7, no. 3, pp. 1–9, 2015.

[6] H. M. Du, S. G. Du, and W. Li, "Probabilistic time series forecasting with deep non-linear state space models," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 1, pp. 3–13, 2023.

[7] Q. Liu, S. Wu, L. Wang, and T. N. Tan, "Predicting the next location: A recurrent model with spatial and temporal contexts," in *Proc. AAAI Conference on Artificial Intelligence*, 2016, pp. 194–200.

[8] H. X. Yao, X. F. Tang, H. Wei, G. J. Zheng, and Z. H. Li, "Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction," in *Proc. AAAI Conference on Artificial Intelligence*, 2019, pp. 5668–5675.

[9] Y. G. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in *Proc. International Conference on Learning Representations*, 2018, pp. 1–16.

[10] S. Y. Li, X. Y. Jin, Y. Xuan, X. Y. Zhou, W. H. Chen, Y. X. Wang, and X. F. Yan, "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," in *Proc. Advances in Neural Information Processing Systems*, 2019, pp. 5243–5253.

[11] S. Fang, Q. Zhang, G. F. Meng, S. M. Xiang, and C. H. Pan, "Gstnet: Global spatial-temporal network for traffic flow prediction," in *Proc. International Joint Conference on Artificial Intelligence*, 2019, pp. 2286–2293.

[12] K. F. Bi, L. X. Xie, H. H. Zhang, X. Chen, and X. T. Gu, "Accurate medium-range global weather forecasting with 3d neural networks," *Nature*, vol. 619, p. 533–538, 2023.

[13] M. Z. Li and Z. X. Zhu, "Spatial-temporal fusion graph neural networks for traffic flow forecasting," in *Proc. AAAI Conference on artificial intelligence*, 2021, pp. 4189–4196.

[14] G. Y. Li, X. F. Wang, G. S. Njoo, S. H. Zhong, G. Chan, C. C. Hung, and W. C. Peng, "A data-driven spatial-temporal graph neural network for docked bike prediction," in *Proc. IEEE International Conference on Data Engineering*, 2022, pp. 713–726.

[15] J. B. Zhang, Y. Zheng, and D. K. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. AAAI Conference on Artificial Intelligence*, 2017, pp. 1655–1661.

[16] L. Zhao, M. Gao, and Z. W. Wang, "St-gsp: Spatial-temporal global semantic representation learning for urban flow prediction," in *Proc. ACM International Conference on Web Search and Data Mining*, 2022, pp. 1443–1451.

[17] J. H. Ji, J. Y. Wang, C. Huang, J. J. Wu, B. R. Xu, Z. H. Wu, J. B. Zhang, and Y. Zheng, "Spatio-temporal self-supervised learning for traffic flow prediction," in *Proc. AAAI Conference on Artificial Intelligence*, 2023, pp. 4356–4364.

[18] S. N. Guo, Y. F. Lin, N. Feng, C. Song, and H. Y. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in *Proc. AAAI Conference on Artificial Intelligence*, 2019, pp. 922–929.

[19] X. Y. Zhang, C. Huang, Y. Xu, L. H. Xia, P. Dai, L. F. Bo, J. B. Zhang, and Y. Zheng, "Traffic flow forecasting with spatial-temporal graph diffusion network," in *Proc. AAAI Conference on Artificial Intelligence*, 2021, pp. 15 008–15 015.

[20] J. Feng, Y. Li, Z. Q. Lin, C. Rong, F. N. Sun, D. S. Guo, and D. P. Jin, "Context-aware spatial-temporal neural network for citywide crowd flow prediction via modeling long-range spatial dependency," *ACM Transactions on Knowledge Discovery from Data*, vol. 16, no. 3, pp. 1–21, 2022.

[21] G. Woo, C. Liu, D. Sahoo, A. Kumar, and S. C. H. Hoi, "Cost: Contrastive learning of disentangled seasonal-trend representations for time series forecasting," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[22] H. L. Ning, X. T. Zheng, X. Q. Lu, and Y. Yuan, "Disentangled representation learning for cross-modal biometric matching," *IEEE Transactions on Multimedia*, vol. 24, pp. 1763–1774, 2021.

[23] J. L. Hu, B. Yang, C. J. Guo, C. S. Jensen, and H. Xiong, "Stochastic origin-destination matrix forecasting using dual-stage graph convolutional, recurrent neural networks," in *Proc. IEEE International Conference on Data Engineering*, 2020, pp. 1417–1428.

[24] Z. R. Xu, Y. B. Wang, M. S. Long, J. M. Wang, and M. KLiss, "Predcnn: Predictive learning with cascade convolutions," in *Proc. International Joint Conference on Artificial Intelligence*, 2018, pp. 2940–2947.

[25] X. Y. Wang, Y. Ma, Y. Q. Wang, W. Jin, X. Wang, J. L. Tang, C. Y. Jia, and J. Yu, "Traffic flow prediction via spatial temporal graph neural network," in *Proc. The Web Conference*, 2020, pp. 1082–1092.

[26] H. Z. Shi, Q. M. Yao, Q. Guo, Y. G. Li, L. Y. Zhang, J. P. Ye, Y. Li, and Y. Liu, "Predicting origin-destination flow via multi-perspective graph convolutional network," in *Proc. IEEE International Conference on Data Engineering*, 2020, pp. 1818–1821.

[27] Z. H. Wu, S. R. Pan, G. D. Long, J. Jiang, X. J. Chang, and C. Q. Zhang, "Connecting the dots: Multivariate time series forecasting with graph neural networks," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 753–763.

[28] H. Y. Chai, S. Y. Tang, J. H. Cui, Y. Ding, B. X. Fang, and Q. Liao, "Improving multi-task stance detection with multi-task interaction network," in *Proc. Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 2990–3000.

[29] R. G. Cirstea, T. Kieu, C. Guo, B. Yang, and S. J. Pan, "Enhancenet: Plugin neural networks for enhancing correlated time series forecasting," in *Proc. IEEE International Conference on Data Engineering*, 2021, pp. 1739–1750.

[30] Q. Q. Fan, M. Jiang, W. T. Huang, and Q. C. Jiang, "Considering spatiotemporal evolutionary information in dynamic multi-objective optimisation," *CAAI Transactions on Intelligence Technology*, vol. 1, no. 1, pp. 1–21, 2023.

[31] S. Y. Lan, Y. T. Ma, W. K. Huang, W. W. Wang, H. Y. Yang, and P. Li, "Dstagnn: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting," in *Proc. International Conference on Machine Learning*, 2022, pp. 11 906–11 917.

[32] R. Cirstea, B. Yang, C. Guo, T. Kieu, and S. Pan, "Towards spatio-temporal aware traffic time series forecasting," in *Proc. IEEE International Conference on Data Engineering*, 2022, pp. 2900–2913.

[33] H. Y. Chai, J. H. Cui, S. Y. Tang, Y. Ding, X. W. Liu, B. X. Fang, and Q. Liao, "Mg-sin: Multigraph sparse interaction network for multitask stance detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 1, no. 1, pp. 1–15, 2023.

[34] R. H. Jiang, Z. N. Wang, J. Yong, P. Jeph, and Q. Chen, "Spatio-temporal meta-graph learning for traffic forecasting," in *Proc. AAAI Conference on Artificial Intelligence*, 2023, pp. 8078–8086.

[35] J. Choi, H. Choi, J. Hwang, and N. Park, "Graph neural controlled differential equations for traffic forecasting," in *Proc. AAAI Conference on Artificial Intelligence*, 2022, pp. 6367–6374.

[36] L. B. Liu, J. J. Zhen, G. B. Li, G. Zhan, Z. C. He, B. W. Du, and L. Lin, "Dynamic spatial-temporal representation learning for traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 11, pp. 7169–7183, 2020.

[37] Y. D. Wang, H. Z. Yin, T. Chen, C. Y. Liu, B. Wang, T. Y. Wo, and J. Xu, "Gallat: A spatiotemporal graph attention network for passenger demand prediction," in *Proc. IEEE International Conference on Data Engineering*, 2021, pp. 2129–2134.

[38] C. P. Zheng, X. L. Fan, C. Wang, and J. Z. Qi, "Gman: A graph multi-attention network for traffic prediction," in *Proc. AAAI Conference on Artificial Intelligence*, 2020, pp. 1234–1241.

[39] Y. Fang, Y. Qin, H. Luo, F. Zhao, B. Xu, L. Zeng, and C. Wang, "When spatio-temporal meet wavelets: Disentangled traffic forecasting via efficient spectral graph attention networks," in *Proc. IEEE International Conference on Data Engineering*, 2023, pp. 517–529.

[40] H. Y. Li, X. Wang, Z. W. Zhang, Z. H. Yuan, H. Li, and W. W. Zhu, "Disentangled contrastive learning on graphs," in *Proc. Advances in Neural Information Processing Systems*, 2021, pp. 21 872–21 884.
[41] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. International Conference on Learning Representations*, 2014, pp. 1–14.
[42] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *Proc. International Conference on Learning Representations*, 2017, pp. 1–22.
[43] A. Gonzalez-Garcia, J. V. D. Weijer, and Y. Bengio, "Image-to-image translation for cross-domain disentanglement," 2018, pp. 1–12.
[44] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," 2016, pp. 1–9.
[45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," 2014, pp. 1–9.
[46] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Proc. International Conference on Machine Learning*, 2018, pp. 531–540.
[47] J. L. Deng, X. S. Chen, R. H. Jiang, X. Song, and I. W. Tsang, "St-norm: Spatial and temporal normalization for multi-variate time series forecasting," in *Proc. ACM International Conference on Knowledge Discovery and Data Mining*, 2021, pp. 269–278.
[48] W. McGill, "Multivariate information transmission," *Transactions of the IRE Professional Group on Information Theory*, vol. 4, no. 4, pp. 93–111, 1954.
[49] S. Srinivasa, "A review on multivariate mutual information," *Univ. of Notre Dame, Notre Dame, Indiana*, vol. 2, pp. 1–6, 2005.
[50] H. J. Hwang, G. H. Kim, S. H. Hong, and K. E. Kim, "Variational interaction information maximization for cross-domain disentanglement," in *Proc. Advances in Neural Information Processing Systems*, 2020, pp. 22 479–22 491.
[51] A. Alemi, I. Fischer, J. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proc. International Conference on Learning Representations*, 2017, pp. 1–19.
[52] L. Z. Han, B. W. Du, L. L. Sun, Y. J. Fu, Y. S. Lv, and H. Xiong, "Dynamic and multi-faceted spatio-temporal deep learning for traffic speed forecasting," in *Proc. ACM International Conference on Knowledge Discovery and Data Mining*, 2021, pp. 547–555.
[53] J. B. Zhang, Y. Zheng, and D. K. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. AAAI Conference on Artificial Intelligence*, 2017, pp. 1–7.
[54] J. Y. Wang, J. W. Jiang, W. J. Jiang, C. Li, and W. X. Zhao, "Libcity: An open library for traffic prediction," in *Proc. International Conference on Advances in Geographic Information Systems*, 2021, p. 145–148.
[55] J. L. Zhang, F. Chen, Y. N. Guo, and X. H. Li, "Multi-graph convolutional network for short-term passenger flow forecasting in urban rail transit," *IET Intelligent Transport Systems*, vol. 14, no. 10, pp. 1210–1217, 2020.
[56] W. Z. Qian, D. L. Zhang, Y. Zhao, K. Zheng, and J. Q. James, "Uncertainty quantification for traffic forecasting: A unified approach," in *Proc. IEEE International Conference on Data Engineering*, 2023, pp. 2992–1004.
[57] L. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, p. 2579–2605, 2009.