Predicting collective human mobility via countering spatiotemporal heterogeneity

Zhengyang Zhou, *Student Member, IEEE,* Kuo Yang, Yuxuan Liang, Binwu Wang, Hongyang Chen, *Senior Member, IEEE,* Yang Wang*, *Senior Member, IEEE.*

Abstract—Human mobility forecasting is the key to energizing considerable mobile computing services. However, we find that the collective mobility suffers the spatiotemporal heterogeneity issue and therefore leads to inferior performances of conventional homogeneous aggregations. Given two fundamental factors, i.e., data and objectives in machine learning, we propose to counter such heterogeneity by improving data utilization and optimization objectives. 1) From data utilization perspective, we discover that such heterogeneity is inherently induced by mobility-related context factors and thus these factors can be exploited to learn heterogeneous mobility patterns. 2) From the optimization perspective, the dependencies among output elements, which give another prior to learning, can extract heterogeneous correlations within output sequences. Specifically, we propose a novel Context-Directional SpatioTemporal Graph Network (CD-STGNet), which tackles the above-mentioned heterogeneity, for achieving accurate mobility predictions. Firstly, we improve data utilization by inputting the encoded context-wise interactions to a direction field learner, which realizes directional spatial aggregations. Secondly, regarding series learning and optimization objectives, a context-trend highway is designed to enable context-aware temporal learning while two regularization objectives are proposed to keep the correlations among predicted elements consistent with the ground-truth. Experiments demonstrate that CD-STGNet surpasses competitive baselines by 13% to 22% and boosts the interpretability of context-directional learning.

Index Terms—Human mobility prediction, spatiotemporal heterogeneity, graph convolutional network, urban computing.

1 Introduction

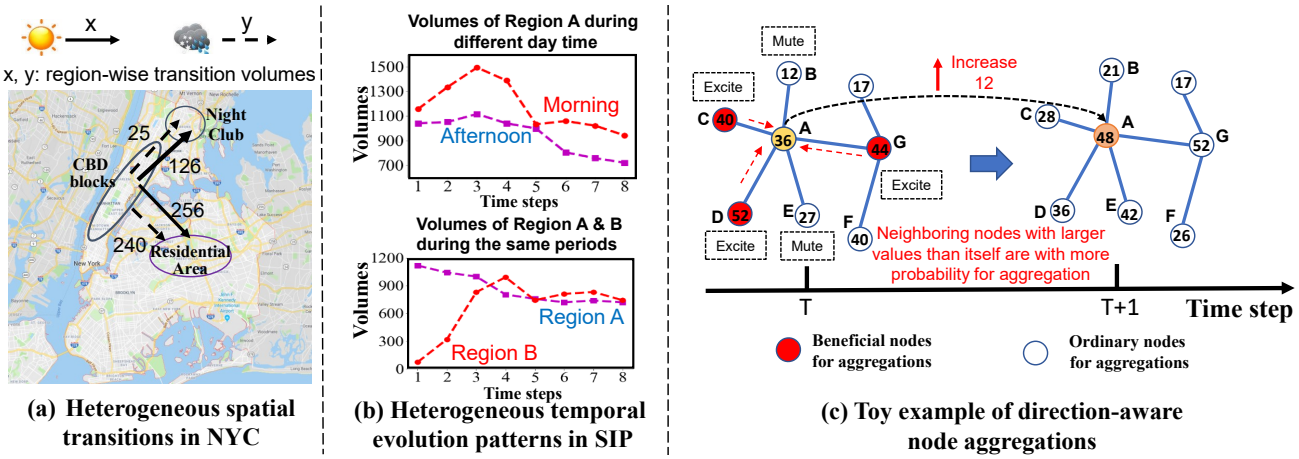
Human mobility, an important index of human-society interactions, offers great potential to diverse mobile computing services, facilitating efficient traffic scheduling [1], [2], urban safety [3], [4], and optimizing data transmission efficiency [5], [6]. Compared to individual mobility, prediction of region-wise collective mobility falls to spatiotemporal learning and can be more valuable for understanding global real-time citywide statuses in urban management and general mobile services. Concretely, administrations can exploit the expected collective mobility to implement road controlling for avoiding the fatal gathering events while ride-sharing platforms can investigate the regularity for taxicab scheduling and order dispatch [7], [8]. In addition, in the Internet of vehicles of open road networks, the mobility prediction of vehicles is an important issue for data transmission scheme to identify the hotspots of signals that contributes to efficient packet exchange [5]. Off-the-shelf collective mobility predictions have achieved promising results from abstracting the regularity of routine-oriented human activities. Technically, the basic idea of existing mobility learning is to build mapping functions from historical observations to future targets by leveraging multi-range [9] and multi-level spatiotemporal correlations [10], [11]. The specific technologies can be classified into Graph Neural Network (GNN) [12], [13] or Convolution Neural Network (CNN) [1] for spatial aggregations, and Recurrent Neural Network (RNN) [14] or Temporal Convolution Network (TCN) blocks [15] for temporal learning. These deep learning-based solutions on mobility forecasting have achieved impressive success.

Actually, we find that the transitions and evolutions of human mobility are prone to be heterogeneous across different temporal steps and spatial urban regions, which significantly impact decisions of urban management and data transmission services. Such phenomenon is formally designated as spatiotemporal heterogeneity in our paper. To this end, the architectures based on homogeneous spatiotemporal aggregations [1], [9], [14], [16], [17] inevitably fall short in capturing their personalized and heterogeneous patterns. In addition, recent research argues that the improvement of forecasting accuracy induced by modifying neural network structures has become incremental, it is of great significance to seek novel perspectives to boost up the learning performances [18], [19], [20]. Therefore, besides neural structures, we propose to explicitly counter the heterogeneity through two additional perspectives, i.e., data utilizations and learning objectives to jointly promote prediction quality where data and objectives are considered as two fundamental factors in machine learning schemes [21].

Firstly, in the data perspective, apart from main mobility observations, we argue that mobility-related covariates that include but not limited to daytime, regional functionality, and weather, can significantly influence the directions of mobility transitions thereby inducing various spatiotemporal patterns of human mobility. More intuitively, the heterogeneity induced by contexts is illustrated by cases

- *Z. Zhou, K Yang, B Wang, and Y. Wang* are with University of Science and Technology of China, China. Y Liang is with Hong Kong University of Science and Technology, Guangzhou, China, and H Chen is with Zhejiang Laboratory, Hangzhou, China.*
  E-mail: {zzy0929, yangkuo, wbw1995}@mail.ustc.edu.cn, yuxliang@outlook.com, dr.h.chen@ieee.org, angyan@ustc.edu.cn.*
- *Prof. Yang Wang is the corresponding author.*

Manuscript received Nov 1st, 2022; revised May 7th, 2023.



(a) Heterogeneous spatial transitions in NYC

(b) Heterogeneous temporal evolution patterns in SIP

(c) Toy example of direction-aware node aggregations

Fig. 1: Examples of spatiotemporal heterogeneity and beneficial node selections for directional aggregations

of New York City (taxicab records in NYC) and Suzhou Industry Park (all-type traffic volumes in SIP). In Figure 1(a), when the weather of NYC transfers from sunny to rainy, there exists a sudden decrease of volume transitions from CBD to night clubs while it shows much stable for volume transitions from CBD to residential areas. In Figure 1(b), two regions of A and B experience completely different trends of volumes at a same temporal period. Based on above, we summarize these covariates as context factors and designate various patterns of spatial transitions and temporal evolutions under diverse contexts as spatiotemporal heterogeneity. In this way, this heterogeneity is attributed to the dynamic and interactive influences of multiple context factors on mobility transitions. With this insight, exploiting context factors to counter the spatiotemporal heterogeneity becomes an opportunity to close the gap between the maximum predictability and existing prediction performances. Fortunately, some recent approaches [15], [22], [23] have taken initial steps on involving contexts into neural networks to fine-tune predictions, and a pioneering work takes contexts as meta factors and makes concatenations with main observations [13] for fine-grained traffic predictions. However, these methods, which directly aggregate contexts with main features, have never considered the interactions of multiple contexts and their explicit impacts on directional aggregations, hence failing to capture the context-induced heterogeneity. Therefore, how to effectively improve the data utilization, i.e., exploiting the contexts to enable dynamically context-aware spatial transitions and temporal trends, still remain unresolved.

Secondly, the learning objective essentially plays significant roles in forecasting models. However, almost all collective mobility learning frameworks [1], [9], [16], [17], [24] only exploit the trivial objectives such as MAPE or MAE to realize the regression, neglecting the potential heterogeneous correlations and dependencies within prediction outcomes. In contrast, serial works have theoretically illustrated that modeling dependencies on element-wise outputs enjoys the benefit of robust prediction by combining Gaussian process and neural process [25], [26]. In a real case of Figure 1(b), each sequence reveals distinctive sequential trends and element-wise dependence, which delivers that the heterogeneity also exists in outputs. Therefore, capturing correlations among elements in predicted sequence can explicitly model the heterogeneity within predicted targets and promote performances for its complements to error-based objective. To this end, in our work, we emphasize the consistency of target-wise dependencies between predictions and groundtruths, to tackle the heterogeneity.

To summarize, given above observations and analysis, these off-the-shelf mobility forecasting solutions fall short in modeling the heterogeneity from two perspectives, i.e., 1) uniform aggregations fail to capture sharp variations and cannot allow the context factors to interpret heterogeneous aggregations, and 2) they never capture the heterogeneity and potential dependencies within predicted outcomes. Considering existing efforts on mobility forecasting and context-aware learning, two sub-challenges remain to counter the spatiotemporal heterogeneity, (i) how to maximally internalize multiple contexts to guide directional message passing and temporal trend modeling, and (ii) how to design informative objectives to capture heterogeneous correlations among predicted target sequences.

In this paper, we first theoretically demonstrate the necessity of context-aware prediction by information theory, which provides a theoretical guarantee to our subsequent solution. After that, we shed light on a two-stage Contextual-Directional SpatioTemporal Graph Network (CD-STGNet) to jointly address above two challenges. In the first stage of CD-STGNet, inspired by the fact that mobility prediction is to explore how mobility transits throughout a city associated with context environments, a Context Directional Spatial Aggregator (CDSA), is designed to predict the primary evolution direction and perform directional message passing with context factors. As illustrated in Figure 1(c), 'increase' is identified to be the predicted primary direction of the central node A, then A's neighboring nodes whose current values larger than itself are considered more beneficial for future target-oriented aggregation, and reasonably highlighted for aggregation. The second stage, Deep Context Temporal Factorization (DCTF), receives the outputs of CDSA and bridges the gap between spatial feature maps and predicted future sequences. More specifically, a context-trend highway capturing mappings

between context factors and predicted sequence is designed, and two novel regularization objectives considering the element-wise correlations and temporal shape-trends are devised. Such highway mapping and novel regularizations respectively enable the heterogeneous evolution modeling and ensure the consistency of element-wise dependencies between groundtruths and predicted sequences. In addition, to cohere multiple objectives and optimize the intractable learning process, we propose an alternate-and-adaptive optimization strategy to alternately train auxiliary and main tasks, and automatically re-weight task-wise importance by leveraging differences of gradient descents. In a nutshell, we make the following contributions.

- This work systematically tackles the collective mobility prediction from the perspective of countering spaptiotemporal heterogeneity, with substantial theoretical, empirical and technical contributions. We first theoretically verify the necessity of context-aware predictions by deriving an inequality regarding the context-conditioned joint entropy, and visualizing three context-aware mobility datasets. Technically, we attribute the spatiotemporal heterogeneity in mobility to context factors and a CD-STGNet is proposed to tackle such heterogeneity.
- We address the spatiotemporal heterogeneity by two aspects, improving data utilization and learning objectives. A node-wise direction learner and context-trend highway by taking contexts as inputs, are designed to respectively perform directional spatial aggregation and context-aware temporal learning. To retain the heterogeneous dependencies among predicted target sequences, two novel objectives of Shape-trend and Covariance, are developed to capture directional temporal trend and heterogeneous node-level correlations.
- We conduct experiments on three different types of human mobility datasets, which further verifies the heterogeneity across different mobility types and demonstrates our CD-STGNet can achieve at least 13% performance gains against competitive baselines. Case studies illustrate the promotion of the interpretability regarding context-induced spatiotemporal heterogeneity.

The rest is organized as follows. Sec 2 formally defines our problem and performs theoretical analysis on the necessity of context-aware prediction. Sec 3 introduces our CD-STGNet for capturing the spatiotemporal heterogeneity, and Sec 4 empirically evaluates our solution. Sec 5 further discusses our insights and method limitations and Sec 6 reviews the related work. Sec 7 finally concludes our work.

2 PRELIMINARY

In this section, we first formally define the basic concepts as well as the problem studied in this work, and provide a theoretical guarantee of our insights and solutions.

2.1 Notations and Problem Definition

Definition 1 (Urban graph and target observations.). *The whole city is discretized into N non-overlapping urban regions and can be constructed as an urban graph $G(\mathcal{V}, \mathcal{E})$, where node set $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$ and edge set $\mathbb{E} = \{e_{ij} \,|1 \leqslant i, j \leqslant N \& \ i \neq j\}$ respectively denote specific urban regions and potential correlations between pair-wise regions. To define fine-grained spatiotemporal observation of human mobility, we divide the time domain into equal intervals $\mathcal{T} = \{1, 2, 3..., T, T+1, ...\}$, and formulate the main observation set as $\{\mathbb{X} = \boldsymbol{X}_{:,t} | t \in \mathcal{T}\}$, where the element $\boldsymbol{X}_{:,t}$ denotes the citywide interval-specific mobility observation while $x_{i,t} \in \boldsymbol{X}_{:,t} (1 \leqslant i \leqslant N)$ represents the specific mobility intensity at region v_i during interval t. The collective mobility can be instantiated as the number of taxi trip records, volumes of all-type traffics or other mobility indicators.*

Definition 2 (Context factors.). *Auxiliary covariates, correlated with main targets but not for predictions, are defined as context factors. Given M types of context factors, the type names of context factors are defined by set $\mathbb{C} = \{C_1, C_2, \cdots, C_M\}$. In our mobility predictions, we can instantiate a three-type context tuple as $\{C_{loc}, C_{ts}, C_w\}$ to represent the factors of location, timestamp, and weather. More specifically, let $\boldsymbol{c}^t = \{c^t_{(m,i)} \in \mathbb{R}^{1 \times d_m} | 1 \leq i \leq N, 1 \leq m \leq M\} (t \in \mathcal{T})$ be the time-varying context observations, where $c^t_{(m,i)}$ is the m-th context descriptor regarding region i and time interval t, and d_m is the dimension of the m-th type context factor.*

Problem 1 (Collective human mobility prediction). *Given historical T-step spatiotemporal observations $\mathbb{X} = \{\boldsymbol{X}_{:,1}, \boldsymbol{X}_{:,2}, ..., \boldsymbol{X}_{:,T}\}$, and the associated time-varying context factors $\boldsymbol{c}^1, \boldsymbol{c}^2, ..., \boldsymbol{c}^T$, we aim to design a function $f(\cdot)$ to counter the spatiotemporal heterogeneity challenge, and perform **collective human mobility prediction** in the following l time steps $\hat{\boldsymbol{Y}}_{:,T+1}, \hat{\boldsymbol{Y}}_{:,T+2}, ..., \hat{\boldsymbol{Y}}_{:,T+l}$ by leveraging context factors and target dependencies.*

$$\{\boldsymbol{X}_{:,1}, \boldsymbol{X}_{:,2}, ..., \boldsymbol{X}_{:,T}; \boldsymbol{c}^1, \boldsymbol{c}^2, ..., \boldsymbol{c}^T\} \xrightarrow{f(\cdot)} \hat{\boldsymbol{Y}}_{:,T+1}, \hat{\boldsymbol{Y}}_{:,T+2}, ..., \hat{\boldsymbol{Y}}_{:,T+l}$$

2.2 Theoretical Analysis for Context-aware Prediction

We theoretically demonstrate that incorporating the context condition can exactly increase the predictability, i.e., the necessity of context-aware prediction, by *information theory*. Given the historical and targeted main observations $\mathbf{X}$ and $\mathbf{Y}$, let us begin with considering one mobility-related covariate, i.e., context factor C_0. The entropy, measuring the degree of chaos in data, can be viewed as the difficulty of fitting corresponding data and thus their predictability. The smaller entropy reflects less discrepancy and larger predictability. Given the specific value $\mathbf{c}_0$ of type C_0, the (x, y) pair will be organized by the category of context factor $\mathbf{c}_0$ and shrink to conditional observations. Correspondingly, since C_0 is the retrieved mobility-related covariate, the regularity of observation pair (x, y) increases due to the filtering of relative context conditions. Let $H(\mathbf{X}), H(\mathbf{X}|\mathbf{c}_0)$ respectively denote as the entropy and conditional entropy of $\mathbf{X}$ and $\mathbf{X}$ conditioned on $\mathbf{c}_0$. We thus have $H(\mathbf{X}) > H(\mathbf{X}|\mathbf{c}_0)$ and $H(\mathbf{Y}) > H(\mathbf{Y}|\mathbf{c}_0)$ according to *condition reduction principle [27].*

We further demonstrate that *the conditional joint entropy also conforms to the condition reduction principle*. As reducing one condition will boost chaos of observation Y, we have

$H(\mathbf{X}|\mathbf{Y},\mathbf{c}_0) < H(\mathbf{X}|\mathbf{Y})$. And given $H(\mathbf{X},\mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}|\mathbf{X})$, we make the following derivation,

$$\begin{aligned} H(\mathbf{X},\mathbf{Y}|\mathbf{c}_0) &= H(\mathbf{Y}|\mathbf{X},\mathbf{c}_0) + H(\mathbf{X}|\mathbf{c}_0) \\ &< H(\mathbf{Y}|\mathbf{X}) + H(\mathbf{X}|\mathbf{c}_0) \\ &< H(\mathbf{Y}|\mathbf{X}) + H(\mathbf{X}) \\ &= H(\mathbf{X},\mathbf{Y}) \end{aligned} \tag{1}$$

It delivers that the joint probability distribution of X and Y can be more regular when they are organized by the context condition $\mathbf{c}_0$, hence increasing the predictability and predictive power from $\mathbf{X}$ to $\mathbf{Y}$ [1]. By assuming M categories of context factors and k_m solid values for each category, the conditional entropy can be extended to the combined condition scenario $H(\mathbf{X},\mathbf{Y}|c_1,c_2,...,c_M) < H(\mathbf{X},\mathbf{Y})$. In detail, there are totally $\prod_{m=1}^{M} k_m$ types of context-wise combinations and the benefits for predictability are determined by the influential correlations of different contexts on $x \to y$ mappings, where the more informative the contexts are, the higher predictability. Therefore, we have verified the necessity of context-aware prediction and provided a theoretical guarantee for our following solutions.

3 METHODS

3.1 Framework Overview

Our mobility prediction solution improves existing ones via countering the spatiotemporal heterogeneity, from two fundamental perspective of machine learning, data and objectives. As illustrated in Figure 2, our Context-Directional Spatiotemporal Graph Network (CD-STGNet) can be decomposed into two cascaded components, i.e., Context Directional Spatial Aggregator (CDSA) accounting for improved context utilizations, and Deep Contextual Temporal Factorization (DCTF) with novel objectives for additional constraints. In CDSA, we first propose a Semantic Context Encoder to capture interactions among multiple context factors. Guided by the direction learner, we generate node-specific direction-aware vector fields for target-oriented beneficial node selection, which realizes the personalized spatial aggregation and outputs the spatial aggregated feature maps. For DCTF, it receives spatial feature maps from CDSA and directly learn the directional transformations from spatial features to multi-step targets. Typically, the context-trend highway instantiates the learnable mappings between context factors and predicted sequence, thus capturing the context-aware latent trends. Besides, the two consistency objectives constraining target-wise sequential shape-trends are introduced to tackle the heterogeneity in predicted outcomes. Finally, an alternate-and-adaptive optimization seamlessly integrates these tasks into a stabilized and adaptive multi-task learning framework.

3.2 Context Directional Spatial Aggregator

Our CDSA aims to efficiently exploit the context factors and perform immediate directional aggregations to tackle spatial heterogeneity. We take the citywide observations regarding the most recent τ time steps away from the nearest prediction step $T+1$ as inputs, i.e. $\mathbf{X}_0 = \boldsymbol{X}_{:,(T-(\tau-1)):T} \in \mathbb{R}^{N\times\tau}$. Recall that the essence of mobility aggregation is to investigate how human mobility spread throughout the city from history to future, and region-specific contextual factors are critical in spatial aggregations. We thus propose our Context Directional Spatial Aggregator, to conduct directional spatial aggregation via generating direction-aware vector fields.

3.2.1 Semantic context encoder

Since various context factors usually interact with each other, building high-quality and interaction-incorporated context embedding is essential to context-aware mobility predictions. However, existing methods for context modeling (e.g., embedding) rarely consider such interactions, leading to a suboptimal modeling of the complex contextual information. To this end, we propose a Semantic Context Encoder to explicitly encode both context-wise self correlations and mutual interactions, transferring discrete context into continuous and interpretable embedding as spatiotemporal aggregation guidance.

We illustrate the architecture of our context encoder in Figure 3. First, for each region i, we concatenate multiple context vectors and impose a learnable correlation weight $\mathbf{w}_{cs} \in \mathbb{R}^{D\times K}$ to obtain self correlation-enhanced representation $(\mathbf{Z}_{cs})_i \in \mathbb{R}^{1\times K}$,

$$(\mathbf{Z}_{cs})_i = \text{Concat}[(c_{m,i}|1 \leqslant m \leqslant M)] * (\mathbf{w}_{cs})_i \tag{2}$$

where K is a hyperparameter controlling the dimension of a semantic context embedding and $D = \sum_{m=1}^{M} d_m$. Second, to capture context-wise interactions, we draw inspiration from relational GCN [28] and take context factors as separate entities to quantify their mutual interaction encapsulated embedding $(\mathbf{Z}_{ca})_i \in \mathbb{R}^{N\times K}$ by,

$$(\mathbf{Z}_{ca})_i = \prod_{m=1}^{M} (\mathbf{c}_{(m,i)} + \mathbf{b}_{c_m})(\mathbf{w}_{ca}^{(m,m+1)})_i \tag{3}$$

The learnable weight $\mathbf{b}_{c_m} \in \mathbb{R}^{1\times d_m}$ helps achieve continuous vectors and $(\mathbf{w}_{ca}^{(m,m+1)})_i \in \mathbb{R}^{d_m\times 1}(m < M)$ captures context-wise interactions, which can be explained as the dynamic context interactions between each context pair. The last weight $\mathbf{w}_{(M,M+1)} \in \mathbb{R}^{d_M\times K}$ is a linear transformation that aligns the context dimension from d_M to K.

Thus, we can parallelly perform citywide context encoding for N urban regions, and finally fuse both self-correlations and context-wise interactions with element-wise additions $\oplus$, resulting in final semantic context representation $\mathbf{Z}_\mathbf{c} \in \mathbb{R}^{N\times K}$,

$$\mathbf{Z}_\mathbf{c} = \mathbf{Z}_{cs} \oplus \mathbf{Z}_{ca} \tag{4}$$

where $(\mathbf{Z}_\mathbf{c})_i, (\mathbf{Z}_{cs})_i, (\mathbf{Z}_{ca})_i$ are the i-th row of $\mathbf{Z}_\mathbf{c}, \mathbf{Z}_{cs}$ and $\mathbf{Z}_{ca}$, respectively. The context factors will be progressively aggregated with learnable weights. Note that it is orthogonal to the permutation of types of context factors, as it conforms to a permutation-invariant learning [29] and semantics can be injected into context-cross weights by task-oriented optimization. Thus, the encapsulated context-wise interactions can potentially counter context-driven spatiotemporal heterogeneity by the subsequent directional spatiotemporal aggregations.

1. The detailed derivation of first line in Eq. (1) can be found in the Appendix.

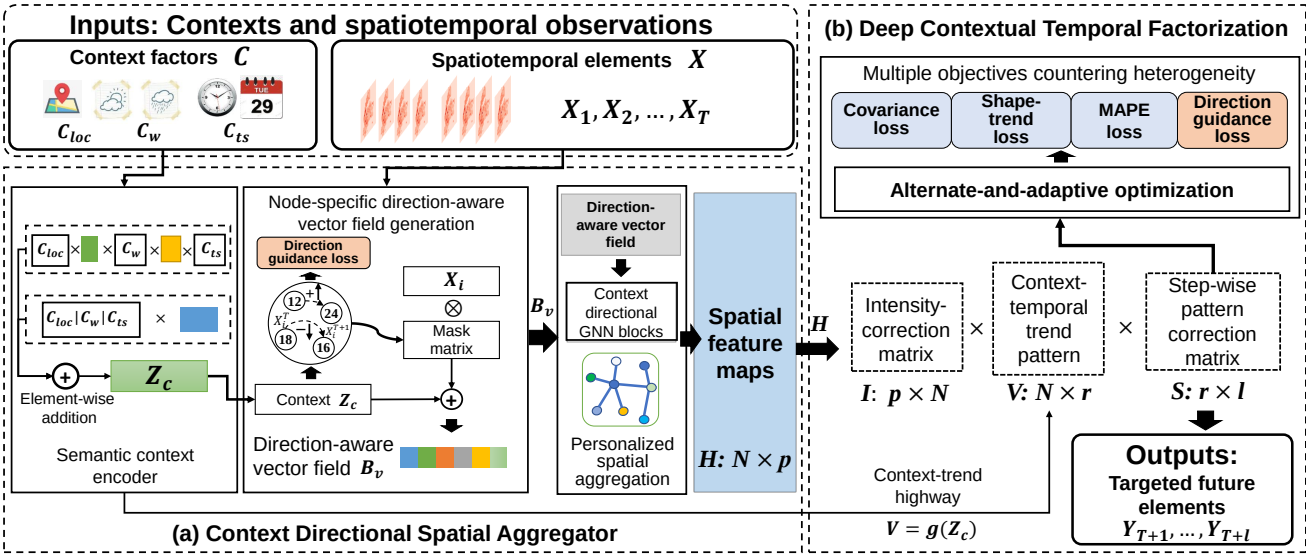


Fig. 2: Context-Directional Spatiotemporal Graph Network. The semantic context encoder and vector field generation respectively receive the context factors and spatiotemporal observations, where the encoder can learn both context-wise interactions and combined context impacts for achieving semantic context representation $\boldsymbol{Z}_c$, while the directional field generation constructs node-specific masking by capturing temporal evolution between adjacent time steps. Noted that the direction guidance takes context representation as inputs and outputs the temporal evolution direction.

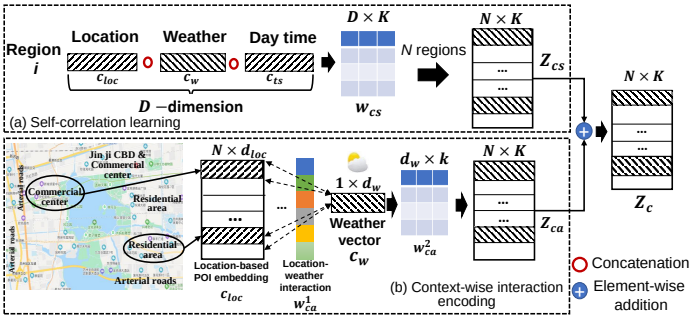


Fig. 3: Architecture of semantic context encoder. The d_{loc} and d_w are dimensions of location and weather vectors.

3.2.2 Direction-aware vector field generation

We have two observations respectively on mobility flows and the spatiotemporal learning models. First, different locations will have various mobility patterns due to context factor-level interactions. And second, the prediction process is to unfold how mobility spread across the city from history to future. Thus, it is necessary to identify node-level personalized aggregation directions by improving utilization of the environmental contexts. Hence, the core idea of our spatial aggregation is to identify the primary evolving direction (from history to future) of each node where the most influential neighbors are aggregated to update the node by virtue of the learned directions. In detail, we borrow the idea of the vector field and flow theory in electromagnetism [30], and propose to generate node-specific direction-aware vector fields. We first define the flow direction of mobility, as the temporal evolving direction of mobility intensity from last historical step T to predicted step $T+1$, and formulate it as,

$$Dir_i = \mathbf{Sign}(x_{i,T+1} - x_{i,T}) \tag{5}$$

where $\mathbf{Sign}(x) = 1$ if $x \geq 0$ and otherwise $\mathbf{Sign}(x) = 0$. This direction, however, cannot be available during the testing phase, thus we design a direction learner to infer the potential variation direction $\widehat{Dir}_i$ for each node i by taking the semantic context encoding as input features,

$$\widehat{Dir}_i = \tanh((\mathbf{Z_c})_i * \mathbf{w}_{dir_i}) \tag{6}$$

We optimize parameters $\mathbf{w}_{dir_i}$ along with $\mathbf{w}_{cs}$ and $\mathbf{w}_{ca}$, $\mathbf{b_{cm}}$ based on the following direction guidance loss,

$$Loss_{dir} = \sum_{i=1}^{N} (\widehat{Dir}_i - Dir_i)^2 \tag{7}$$

With this direction objective, we are capable of capturing the mapping regularity from combined context embedding to the node-wise evolving direction. After that, we can exploit a direction learner to generate node-specific direction-aware vector fields and further select beneficial nodes towards expected aggregation. For a specific node, we consider its neighboring nodes those value differences (between neighbors and itself) are with the same direction of its own temporal evolution as the beneficial nodes for aggregation. We further impose a masking mechanism to filter out those non-beneficial nodes, which are the complementary set to beneficial ones. As intuitively shown in Figure 1(c), the direction learner first predicts the directions of input-to-target variations, and CDSA calculates the differences of values between neighbors and node itself. Actually, we are expected to highlight the variation consistency between neighbors and next-step observation, and filters out node values those are at opposite directions from the predicted targets. Recall that product of numbers with the same sign and opposite signs can respectively contribute to positive and negative values. For each time step, we formulate

the pair-wise masking element in $\mathbf{AM}$ with an indicator function $\mathbf{Ind}(\cdot)$ and max-element selection,

$$(\mathbf{AM})_{\mathbf{ij}} = \max\{\mathbf{Ind}(\mathbf{Sign}(x_{j,T} - x_{i,T}) \cdot \widehat{Dir_i}), \varepsilon_j\} \quad (8)$$

where $j \in \{1, 2, ..., N\}$ and $\varepsilon_j < 1$ is a random noise approaching 0, which allows a few chances to aggregate implicit correlated nodes. $\mathbf{Ind}(x)$ is an indicator function that $\mathbf{Ind}(x) = 1$ when $x > 0$, otherwise $\mathbf{Ind}(x) = 0$. If the neighbor-self difference and next-step self variation are with the same sign, **Ind** will keep as 1 thus we can utilize $'\max'$ to simultaneously excite the beneficial nodes and mute the trend-inconsistent nodes for aggregation. Therefore, the additional benefit of direction-aware mechanism is that when there exists a potential sharp trend, the immediate sharp change can be efficiently captured and preserved with our direction learner and trend-consistent nodes aggregations.

So far, we take the learned variation direction and directional masking mechanism to generate directional vector fields for each node. Considering contexts and node variation directions, we take the most recent observations as input and deactivate less correlated elements. To compensate for the sparsity of above masked direction matrix, we also impose a linear transformation on context encodings to establish explicit context-involved adjacency for context-aware neighbor selection. We then combine these two components to generate final direction-aware vector fields $(\mathbf{B}_v) \in \mathbb{R}^{N \times N}$,

$$(\mathbf{B}_v)_i = ((\mathbf{X_0}_{:,-1})^T \odot \mathbf{AM}_i + (\mathbf{Z_c})_i(\mathbf{W_c})_i) \odot (\mathbf{W_v})_i \quad (9)$$

We denote $\odot$ as Hadamard product, while $(\mathbf{W_c})_i \in \mathbb{R}^{K \times N}$ and $(\mathbf{W_v})_i \in \mathbb{R}^{1 \times N}$ are two learnable transformations to achieve N-dimensional direction field embedding.

3.2.3 Context directional graph neural network

Given the directional vector fields $\mathbf{B}_v$, we can construct our context directional spatial aggregator by feeding the last τ-step historical observations $\mathbf{X}_0$, and stacking a two-layer graph neural network,

$$\mathbf{H} = \mathbf{B}_v(\mathbf{B}_v\mathbf{X}_0\mathbf{W}_s^0)\mathbf{W}_s^1 \quad (10)$$

where $\mathbf{H} \in \mathbb{R}^{N \times p}$ is the aggregated spatial feature maps, and $\mathbf{W}_s^0 \in \mathbb{R}^{\tau \times q}$, $\mathbf{W}_s^1 \in \mathbb{R}^{q \times p}$ are two learnable parameters for aligning feature dimensions.

Distinctions. We distinguish our context-aware direction learner from attention layers [12] and other adaptive adjacent matrix-based spatiotemporal learning [17], [24], [31], [32] as follows. First, the attention layers cannot identify the evolving direction of each node with an explicit criterion. Second, we advance the adaptive adjacent matrix generation in previous studies, which were usually generated by matrix multiplication, towards a directional and context factor-guided manner. This goal is achieved by exploiting context factors and considering node-specific temporal evolving direction as intermediate learnable objectives for guidance, contributing to the improved data utilization for mobility forecasting.

3.3 Deep Contextual Temporal Factorization

Instead of utilizing traditional sequential learning techniques that suffer recursive computations and temporal error accumulation (e.g., RNN-based and Convolution-based methods), inspired by AGCRN [24], we propose a Deep Contextual Temporal Factorization (DCTF) to directly make projections from spatial feature maps to multi-step predictions. To endow the learning capacity, DCTF factorizes three learnable transformations to respectively learn the intensity corrections, context-trend patterns and step-wise transformation. In DCTF, the context-trend highway enables our framework to capture the context-induced latent trend patterns. We specifically introduce a novel element-wise shape-trend objective to model the temporal trends of output sequences thus constraining their intra- and inter-sequence correlation consistent with groundtruth. Further, an alternate-and-adaptive parameter optimization strategy is proposed to seamlessly cohere multiple learning objectives and resolve the issues of both task entanglement and task-wise importance weighting.

3.3.1 Context-assisted temporal factorization

To accommodate the context-related patterns, we factorize the spatial-to-temporal mappings into three learnable transformations, regarding intensity correction, context-trend modeling and step transformations where the multiple groups of trainable parameters are dedicated to improve the fitting capacity. Given aggregated spatial feature maps $\mathbf{H}$, we can obtain the predicted spatiotemporal results $\widehat{\mathbf{Y}}$ by,

$$\widehat{\mathbf{Y}} = \mathbf{H} * \mathbf{I} * \mathbf{V} * \mathbf{S} \quad (11)$$

To be specific, since the learnable matrix $\mathbf{I} \in \mathbb{R}^{p*N}$ is aimed at correcting intensity from spatial features, thus it is initialized with an identity matrix $\mathbf{E}$ with Gaussian noise κ, i.e., $\mathbf{I}$=$\mathbf{E}$ + κ where $\kappa \sim N(0, \delta)$ [2], and it can be trained in a position-sensitive manner. This small modification allows the learnable $\mathbf{I}$ to be easily trained with final objectives. To explicitly incorporate the combined context conditions, the trend pattern matrix $\mathbf{V} \in \mathbb{R}^{N*r}$ is designed by establishing the highway between combined context and the latent compressed temporal trends, where the compressed trend represents a customized latent transformation, and r is the dimension for the latent pattern. In this way, we formulate $\mathbf{V}$ as a function of semantic context encoding $(\mathbf{Z_c})_i$ by,

$$\mathbf{V}[(\mathbf{Z}_c)_i] = g((\mathbf{Z_c})_i) = (\mathbf{Z_c})_i\mathbf{w}_{tr}^i \quad (12)$$

where $\mathbf{w}_{tr}^i \in \mathbb{R}^{K \times l}$ is the learnable weight for capturing the sequential trend regarding node i. Finally, the step-wise pattern matrix $\mathbf{S} \in \mathbb{R}^{r*l}$ is to directly transfer the latent compressed context-trend representation to targeted l steps, where we constrain $r, l \ll N$.

Although the temporal learning process seems non-trivial, three strategies above essentially guarantee the learning quality. First, the total parameters for in DCTF is $O(pN + Nr + lr) \approx O(\lambda N)$ with $\lambda = p + r$, acceptable for our training. Second, the modifications of the noise incorporated identity matrix allows for tiny corrections in our training meanwhile retaining the main patterns. Thirdly, the

2. δ is set to 0.1 according to repetitive experiments.

context conditioned temporal trend learning can enhance the context guidance of parameter learning.

Eventually, we denote element $\widehat{Y}_i(j)$ in $\widehat{\mathbf{Y}}$ as the predicted spatiotemporal element at i-th node during interval j, and so does $Y_i(j)$. More specific objectives that further enhance temporal trend modeling and address the spatiotemporal heterogeneity will be dissected as follows.

3.3.2 Multiple objectives for target-wise dependency learning

Existing methods usually capture spatiotemporal correlations from input feature perspectives, ignoring potential dependencies among targets. Given the opportunity that explicitly constraining consistency of element-wise dependencies among targets can imitate conditional neural process thus enabling the neural network to be more robust to noise [25], [26], designing novel objectives modeling target-wise dependencies can potentially elevate prediction performances where it is further empirically demonstrated in ablation studies. From a spatial-temporal perspective, we decompose such dependencies into variable-wise spatial correlations and temporal trend variations.

Covariance objective. Since covariance measures how each pair of variables moves together from their mean, and reflects variable-wise dependencies, we employ Covariance to calculate node-wise spatial correlations by,

$$\sigma_{ij} = (\mathbf{Y}_i - \overline{Y}_i)(\mathbf{Y}_j - \overline{Y}_j)^T \tag{13}$$

where $\mathbf{Y}_i$ is a sequence-shape vector of targeted i-th node sequence, while $\overline{Y}_i$ is a scalar value denoting mean of $\mathbf{Y}_i$. We impose the dot product of two difference sequences to achieve σ_{ij}, which can be viewed as the proximity between nodes i and j, corresponding to reflecting pair-wise spatial dependencies. Similarly, we can perform the same calculation to obtain covariance of predicted targets $\widehat{\sigma_{ij}}$. To maintain the consistency of target-wise spatial dependencies between predictions and groundtruth, we minimize the MSE between these two equally computed items, realizing our Spatial Covariance Loss,

$$Loss_{\text{cov}} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\sigma_{ij} - \widehat{\sigma}_{ij})^2 \tag{14}$$

Shape-trend objective. The directions and intensities of element-wise variations can characterize the overall shape and sequential trend for each sequence. We then propose a shape-trend vector, computed by the difference of consecutive element values among targets, to capture temporal trend variation directions. Each element in a trend vector $\widehat{\mathbf{Y}}_i(j)_{diff}$ can be calculated by,

$$\widehat{Y}_i(j)_{diff} = \widehat{Y}_i(j) - \widehat{Y}_i(j-1), 1 < j < l \tag{15}$$

Similarly, we can obtain the shape-trend vector sequence of groundtruth $Y_i(j)_{diff}$. As cosine similarity is direction-oriented and intensity-agnostic, we thus minimize their sequence-wise cosine similarities to constrain the trend consistency between predictions and groundtruth,

$$Loss_{tr}(Y_{diff}, \widehat{Y}_{diff}) = \frac{1}{N}\sum_{i=1}^{N}\exp\left(-\frac{(Y_i)_{diff}(\widehat{Y_i})_{diff}}{||(Y_i)_{diff}||||(\widehat{Y_i})_{diff}||}\right) \tag{16}$$

We also integrate the traditional element-wise MAPE loss $Loss_{mape}$ into our multiple objectives to encourage the intensity consistency, and finally achieve our multi-objective learning. To enhance the intuitive understanding of our multi-objective losses, we illustrate the execution process of our spatial covariance and shape-trend losses in Figure 4.

In fact, these two novel constraints can be complementary to traditional error-based objectives to tackle target-wise heterogeneity. For covariance loss, the covariance itself indicates node-wise inter-correlations by product operation when $i \neq j$, while it represents intra-node self-correlations when $i = j$, contributing to an imitated CNP for robustness. And the shape-trend loss can consistently constrain the temporal trend similarity between predicted targets and groundtruth, resulting in a directional temporal trend learning in our CD-STGNet.

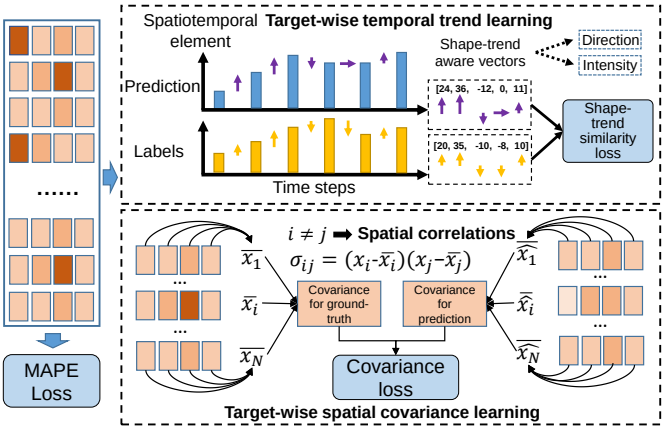


Fig. 4: Multiple objectives constraining target-wise dependencies. The losses constitutes of total objectives are highlighted in three blue boxes.

By assuming all learnable parameters as Θ and combining both multi-objective deep factorization and intermediate direction guidance, we arrive our final integrated objectives,

$$Loss_{MO}(\Theta) = \gamma_1 Loss_{mape} + \gamma_2 Loss_{tr} + \gamma_3 Loss_{\text{cov}} + \gamma_4 Loss_{dir} \tag{17}$$

where $\gamma_i (i = 1, 2, 3, 4)$ are four weighting parameters balancing the losses among different objectives.

Distinctions. Our work can be an improved version of the decoders in existing literature, i.e., fully-connected layers for temporal prediction, but with two dedicated modifications. One is the context-trend highway that endows the framework to learn context-specific latent temporal patterns of each node. And the other is the two novel objectives that preserve the heterogeneous node-wise spatiotemporal correlations among target-wise elements, not only realizing the directional trend modeling but also possessing the property of anti-noise in predictions. These objective-based modifications can be deemed as an improved contribution to accurate and robust models from a fundamental machine learning view.

3.3.3 Alternate-adaptive optimization strategy

Given four tasks and corresponding task-wise weightings, the issues of task disentanglement and task-wise importance optimization can be intricate and non-trivial, as the

loss functions of different tasks usually experience various patterns during optimization process while performing grid searching is highly expensive. To cohere these multiple tasks smoothly and tackle the intractable optimization challenges, we devise an alternate-and-adaptive optimization strategy. The intuition is that the optimization process can be decomposed by different stages by freezing specified tasks and parameters, and the task-wise importance can be balanced by the relative gradient difference among total descents. Particularly, since the direction learner for vector field generation can be separated from main tasks, inspired by coordinate gradient descent [33], we introduce an alternate training strategy. It imposes optimization on direction learner only with four groups of parameters $\mathbf{w}_{dir_i}$, $\mathbf{w}_{cs}$, $\mathbf{w}_{ca}$, $\mathbf{b}_{\mathbf{c_m}}$ by freezing other parameters in the first stage, which intends to achieve a satisfactory direction learner and vector field-based aggregation. And then we train the full parameters in the second stage, simultaneously fine-tuning the above direction learner-specific parameters. This process can be implemented alternately and further allows a well-learned direction guidance and semantic context embedding to generate vector fields.

Secondly, to tackle the non-trivial task-wise weighting issue during joint learning process, we devise an adaptive weight updating strategy. It is formulated by assigning subscripts $1 \sim 4$ to above four losses, representing MAPE, shape-trend, covariance and direction losses. We initialize each weight as the proportion of corresponding absolute loss value to the their summations,

$$\gamma_i = \frac{|Loss_i|}{\sum_i |Loss_i|} \tag{18}$$

Then we can dynamically re-weight the losses according to the task-specific relative gradient difference among total four descents at last epoch,

$$\gamma_i^{'} = \frac{\nabla_\theta Loss_i}{\sum_i \nabla_\theta Loss_i} \tag{19}$$

In this way, all learnable parameters in our CD-STGNet can be optimized alternately and adaptively. We apply Adam as the gradient optimizer for our task [34].

3.4 Complexity Analysis of CD-STGNet

We analyze the time complexity of training our framework. Since the context encoding dimension K and context-wise embedding d_m satisfy $d_m, K \ll N$, our training complexity becomes the summation of total trainable parameters of $(\mathbf{W_c})_i \in \mathbb{R}^{K*N}, (\mathbf{W_v})_i \in \mathbb{R}^{1*N}, I \in \mathbb{R}^{p*N}, V \in \mathbb{R}^{N*r}, S \in \mathbb{R}^{r*l}$. Hence, it can be simplified to $O(\lambda N)$, which is linear times of node number N and acceptable for training.

4 EXPERIMENT

We collect three types of human mobility datasets and their contextual factors. We perform diverse experiments on baseline comparisons, ablative studies, visualized case studies to verify the intuition of our context-induced heterogeneity and our context-guided aggregations.

TABLE 1: Dataset statistics (m: million, k: thousand)

Dataset	Category of datasets	# of records	Time Span	# of regions
SIP	Surveillance	2.7 m	01/01/2017-03/31/2017	108
	Weather	4.3k		
NYC	Taxi trips	7.5 m	01/01/2017-05/31/2017	354
	Weather	7.4k		
Metr-LA	Loop detectors	4.9 m	03/01/2012-06/30/2012	207
	Weather	5.7k		

4.1 Dataset Description

We collect three types of real-world human mobility datasets from perspectives of all-type traffic surveillance volumes of Suzhou Industry Park (SIP), taxicab trip volumes of NYC (NYC)[3], and highway mobility of Metr-LA loop detectors (Metr-LA)[4]. The statistical descriptions of datasets are figured in Table 1. Particularly, to perform context-directional learning, we incorporate multiple context factors, i.e., region descriptions, timestamps (day of week, hour of day and holiday indicators), numerical weathers (weather categories, precipitation, dew, visibility, et, al)[5] associated with corresponding spatiotemporal scopes. Therefore, our CD-STGNet is not only capable of capturing both periodicity and closeness patterns, but also explicitly encourages more efficient utilization of these widely available contexts.

4.2 Implementation Details

For each dataset, we organize them as groups of samples and divide them into 60%, 30% and 10% for training, testing and validation, respectively. We encode the categorical context with one-hot embeddings and transfer them into fixed-length vectors. Our target is to predict the mobility of the next 6 slots based on the current 12 frames ($\tau = 12$) following empirical settings [9], [15], [17]. Each baseline and our network are implemented five times and the averaged errors are reported.

Regarding implementation of context factor learning, we encode the numerical weather into fixed-length vectors $c_{(w,i)} \in \mathbb{R}^{1\times d_w}$, transfer the day intervals and day types into integer values and compress them into a vector $c_{(ts,i)} \in \mathbb{R}^{1\times d_{ts}}$. The implementation of semantic context encoder is three-fold. First, we encode the numerical weather into fixed-length vectors $c_{(w,i)} \in \mathbb{R}^{1\times d_w}$, and secondly, we transfer the day intervals and day types into integer values, where vector $c_{(ts,i)} \in \mathbb{R}^{1\times d_{ts}}$. Third, we initialize random embedding $c_{(loc,i)} \in \mathbb{R}^{1\times d_{loc}}$ for each region to capture personalized location-based patterns. In our work, these three types of contexts will be fed into the semantic encoder and achieve the context-wise interaction learning. For baselines, we directly feed the concatenated context embedding into placeholders of baselines if they have. Noted that the urban graph in our solution is a virtual topology that is derived by the adjacent matrix and the evolving direction learner.

3. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
4. https://github.com/liyaguang/DCRNN
5. Collected from API: https://api.weather.com

For general settings, all the methods are implemented using Tensorflow 1.14.0 or Pytorch 1.10.0 and evaluated on one Tesla V100 GPU. To guarantee fair comparison, we perform grid search to tune the hyperparameters for all baselines over the three datasets. For ours, we stack 3 GCN layers on SIP and Metr-LA, and 2 GCN layers on NYC, and set 1 LSTM layer across all datasets. More specifically, for SIP, we instantiate location dimension $d_{loc} = 16$, context embedding $K = 64$, for NYC, $d_{loc} = 6, K = 8$ and for Metr-LA, $d_{loc} = 32, K = 32$. The initial learning rate is set to 0.0001 with an 0.98 attenuation rate every 10 epochs.

4.3 Performance Comparisons

4.3.1 Baselines

We categorize baselines into context-incorporated and context-agnostic solutions. All baselines except MDL follow the setting of 12 step inputting while MDL follows the setting of three segments of closeness, periodicity and trend. For each baseline, we explicitly provide the loss functions they utilized for further discussion. Besides, to test the model generalization and pluggability, we also modify our network by plugging other GNN-based solutions.

(1) Context-incorporated Baselines:

- **Traffic transformer** is inspired from Google's Transformer and proposed to capture the continuity and periodicity of time series, as well as the spatial dependency [35].
- **STG2Seq** is a hierarchical graph convolution for taxicab passenger demand prediction, considering context factor fusion with element-wise additions [15].
- **ST-SSL** is a spatiotemporal forecasting solution from contrastive learning perspective, which also injects geographical contexts into regional embedding [36].
- **CD-STGNet+GraphWave** is a variant of our approach which replaces our GNN with GraphWaveNet as a variant [37].
- **CD-STGNet+MixHop** is a variant of our approach which replaces our GNN with an advanced MixHop that leverages weighted layer fusion [38].

(2) Context-agnostic Baselines:

- **HA** refers to historical averaging solution for sequence forecasting.
- **AGCRN** is a combination of adaptive GCN and GRU for spatiotemporal forecasting [24].
- **STFGNN** is a GNN-based framework that jointly learns localized heterogeneity and global homogeneity with data-driven graph generation [16].
- **MDL** is a state-of-the-art collective human mobility forecasting method, which is inherited from ST-ResNet [1] and simultaneously models nodes and edges with multiple deep learning tasks. [39].
- **Graph-WaveNet**, is an improved version of DCRNN [14], which designs a learnable dynamic region-wise proximity and dilated convolutions for spatial-temporal learning [37].
- **MTGNN** couples GNN and temporal convolutions to capture underlying series-wise spatial-temporal dependencies in implicit graph structures [17].
- **MRA-BGCN** jointly learns edge-wise and node-wise graphs with a bicomponent graph and multi-range attention [9].

4.3.2 Analysis of performances against competitors

The averaged numerical error results are reported in Table 2. Our framework outperforms the best competitor by 13.52% (traffic transformer), 22.92% (STG2Seq) on SIP and NYC respectively, and achieves comparable performance as traffic transformer on well-studied dataset Metr-LA.

Context-aware and context-agnostic solution comparisons. Although STG2Seq and traffic transformer have incorporated timestamps and weather contexts into learning schemes based on MLP, they still fail to exploit the guidance role of context, resulting in unsatisfactory results. In contrast, thanks to context-aware adaptive topology and target-wise correlation constraints, our CD-STGNet achieves exciting performances almost on all datasets for next 6-step predictions. To test the generality of our framework, we replace the GNN in our CD-STGNet with another two popular graph neural modules, GraphWaveNet and MixHop, for addressing our learning tasks. The superior results of these two modifications to other baselines can verify the generality of our well-designed direction learner and temporal learning modules. The context-agnostic solutions perform relatively inferior to context-aware methods on SIP and NYC, while the seemingly reasonable results haven't achieved on Metr-LA. We ascribe this phenomenon to that different datasets may be suitable for distinctive fusion mechanisms which will be elaborated in followings sections.

Different performances across baselines and loss functions. Another interesting observation is that baselines may perform diversely across different datasets since these solutions may be designed based on specific tasks and datasets. Specifically, MRA-BGCN performs worse than other baselines on SIP and NYC, due to this model may highly depend on the predefined graph structure, while there is no such groundtruth on these two datasets. And STG2Seq, which is tailored for taxi demand prediction, only obtains the best results on NYC trip records. Besides, noted that our CD-STGNet achieves much more superior results than MDL on Metr-LA/SIP and comparable performances on NYC. The reason may lie in that 1) Metr-LA with less variations are more prone to be affected by closeness observations rather than weekly/daily periodicity, 2) our context-aware strategy exactly makes sense. Thus, the heterogeneous performance improvements across datasets can be attributed to the model sensitivity to different data properties. Furthermore, regarding loss functions, we find that even though our integrated loss can prominently improve performances, the types of loss functions do not play significant roles during learning process but the designed auxiliary tasks and additional data utilization do, e.g. transformer-based sequence learning and contextual factor incorporation. Actually, the learning tasks and covariate datasets separated from main targets can be viewed as regularization and complementary prior to main tasks, which leads to a superior optimization. Therefore, benefiting from the exploiting of context factors for directional spatial aggregation and temporal trend learning, our CD-STGNet achieves superior performances over almost all baselines.

TABLE 2: Performance comparisons on three datasets. The best results are in bold and the second best are underlined.

	Loss function	SIP			NYC			Metr-LA		
		MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE
Traffic Transformer	MAE	37.76	23.29%	83.50	32.19	34.67%	76.64	3.52	9.71%	6.33
STG2Seq	MSE	64.80	31.66%	144.23	11.08	14.61%	22.16	5.86	22.20%	10.87
ST-SSL	MAE	90.30	25.10%	170.98	15.38	25.29%	32.73	4.14	22.34%	8.13
HA	MAE	81.24	40.25%	150.23	16.12	33.14%	34.41	6.49	26.41%	12.74
AGCRN	MAE	57.74	25.37%	120.76	11.72	15.85%	24.94	5.88	21.55%	11.54
STFGNN	Huber Loss	36.60	24.45%	84.96	11.76	15.50%	24.98	3.34	10.67%	6.55
MDL	RMSE	82.56	34.55%	192.40	12.48	18.62%	27.62	3.27	36.67%	6.52
GraphWaveNet	MAE	95.96	45.91%	218.11	14.45	27.39%	30.68	3.82	12.06%	7.49
MTGNN	MAE with L2	41.71	32.61%	92.28	16.27	30.98%	34.63	3.14	**8.19%**	6.17
MRA-BGCN	MAE	92.48	42.20%	203.73	17.85	34.23%	36.25	3.19	8.32%	6.27
CD-STGNet	Hybrid Loss	33.17	20.14%	74.01	11.36	**11.26%**	**22.07**	3.13	8.32%	**6.15**
CD-STGNet+WaveNet		36.12	22.00%	80.26	11.94	18.60%	25.42	**3.12**	10.50%	6.22
CD-STGNet+MixHop		**32.60**	**19.80%**	**72.45**	**10.88**	13.40%	23.16	3.28	11.15%	6.45

TABLE 3: Performances on ablative spatiotemporal learning

Variants	MAPE		
	SIP	NYC	Metr-LA
CDSTG-SC	21.48%	23.12%	14.92%
CDSTG-DL	21.22%	18.45%	14.73%
CDSTG-ConIn	21.85%	19.23%	15.06%
CDSTG-CT	24.62%	17.22%	11.48%
CDSTG-SDT	26.28%	15.20%	12.56%
CDSTG-TDT	24.81%	17.44%	14.10%
CDSTG-AdaWeg	26.47%	17.32%	11.25%
CDSTGNet	**20.14%**	**11.26%**	**8.32%**

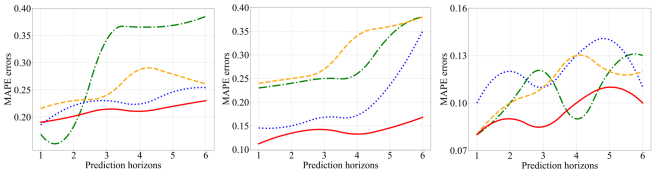


Fig. 5: Comparisons on multi-step prediction performances

Multi-step prediction performances. Since we aim at predicting sequential mobility, the citywide forecasting performances of next 6 steps (average the MAPE errors to the city level) are illustrated in Figure 5, by comparing ours with several selected high-quality baselines. In particular, MTGNN are observed to be limited in predicting only two or four steps on SIP and NYC, respectively, leading to lower overall performances than ours. Traffic transformer and STFGNN, which are dedicated for traffic forecasting, achieve barely satisfactory performances due to the context-encoded mechanism and modified DTW for sequential pattern extraction. Also promisingly, our CD-STGNet surpasses all of them with a significant margin, especially on farther horizons, which verifies the success of our novel objectives in constraining the trend and target-wise consistency between groundtruth and predictions.

Finally, even diverse performances, we can still conclude that the context-directional learning not only promotes the data utilization, also leads to improved performance with learnable and directional spatiotemporal aggregations.

4.4 Ablative Study

To test the effectiveness of each component, we successively remove following components or replace them with ordinary modules as ablative variants,

(1) CDSTG-SC: Remove the Semantic Context encoder and utilize original context embedding for prediction. **(2) CDSTG-DL**: Remove the direction learner loss. **(3) CDSTG-ConIn**: Remove the context-involved adjacency in $\mathbf{B}_v$. We separate the contributions of context-involved adjacency and the directional masked matrix AM to test the necessity of context-based adjacency design in $\mathbf{B}_v$. **(4) CDSTG-CT**: Replace the Context-Trend highway with plain fully-connected layers. **(5) CDSTG-SDT**: Remove the objective of spatial covariance of targets. **(6) CDSTG-TDT**: Remove the objective of shape-trend dependence of targets. **(7) CDSTG-AdaWeg**: Replace the adaptive weighting scheme in multi-task learning with equal weights.

Ablation results demonstration. The quantitative results are shown in Table 3. As illustrated, the guidance of contextual factors consistently plays the most significant role in performance gains. Specifically, the spatial dependency of target-wise constraints, semantic context encoding and adaptive weights respectively play the most significant role on three types of datasets. This observation also manifests that homogeneous aggregations without direction guidance can lead to suboptimal performances and further verify the rationality of our solution. The performance of CD-STGNet-ConIn has a slight downtrend, which indicates that the context-involved aggregation can exactly compensate for the sparsity of vector field matrix to enhance performance. Furthermore, with the adaptive multi-task weighting, we observe a more stable and faster convergence rate as well as improved performance. Numerically, on SIP, our CD-STGNet gets convergence at the epoch of 70 on non-adaptive optimization while obtains convergence at 40 epoch on the adaptive one. To sum up, we have corroborated the effectiveness of the intuitions and designs of above six component designs.

Module sensitivity across cities. We further observe that different cities are with heterogeneous sensitivity to different sub-modules in our CD-STGNet. The reason can be summarized as that, from the data science view, the property of dataset (nature of the city) dominates how models and different modules impact the predictability of corresponding data, i.e., the performance quality. Specifically, the property can be diverse, such as the dependence degree of context factors, the degree of data fluctuations, the consistent correlations between nodes, and so on. For instance, the semantic context embedding and Context-involved adjacency can be more friendly to datasets which are more dependent on contexts, while the spatial covariance module tends to enable more improvements on the datasets with more causal node-level correlations. Therefore, the potential reason of higher improvements with Spatial Covariance on SIP may be the more consistent node-wise correlation, while the more dependence of contexts on NYC may contribute to better performances with semantic context embedding and context-involved adjacency in $\mathbf{B_v}$.

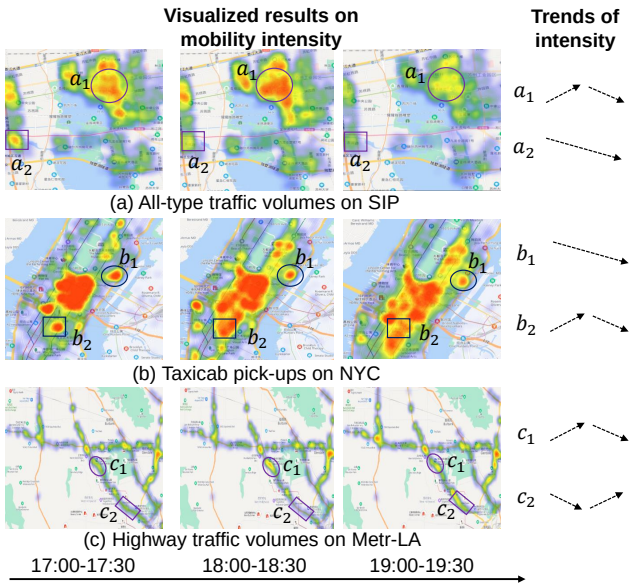


Fig. 6: Various patterns of mobility types

4.5 Case Study

In this section, we will demonstrate some real-world cases and prediction results, to verify the rationality of our intuitions and solutions from diverse perspectives.

Various patterns of mobility types. We visualize the mobility density of three consecutive steps on three datasets in Figure 6, where each set represents one typical mobility type. We also highlight the trends of two selected regions on each set to enhance an intuitive comparison. It is observed that 1) For each set, different regions reveal non-similar patterns for their specific functionalities and interactions between contexts, 2) The heterogeneity across various mobility types is indicated, e.g., taxicabs tend to be concentrated on downtown while all-type vehicle volumes will be more evenly and widely distributed throughout the whole city. This observation can inspire vehicle-aware management for urban traffics and context-aware urban data transmission schemes among vehicles. 3) The regional mobility intensities evolve dynamically and diversely, thus the diversity can further interprets the necessity of directional spatial-and-temporal aggregation in a qualitative and intuitive manner.

Visualization of context embedding. Figure 7 (a) demonstrates the learned adjacent matrix of one step on SIP. In subfigure (b)(c), we select three representative frames of context embedding (learned with the guidance of direction learner) on near steps of above adjacent matrix. As observed, due to daily commutes and human routines, context embedding during morning share similar patterns with those of evening while they show fully different patterns from noon hours, verifying the dynamics and heterogeneous mappings from context to spatiotemporal targets. In this way, the node-specific and context-aware patterns are exactly captured in the context embedding with our target-oriented objectives and directional aggregation filters.

Visualization of predicted sequences. The predicted sequences of two correlated nodes (68, 73) are shown in Figure 7 (d), and we also visualize the predicted sequences of our solution as well as the best baseline (Traffic Transformer) on SIP for comparison. Three observations are achieved. 1) The correlations among neighboring nodes can be well-learned in the adjacent matrix. 2) Our solution makes better initialization and obtains narrower margins between groundtruth and predictions than the transformer-based model does, where the improved performances can be attributed to our direction designer and target-wise directional temporal trend objectives. 3) Previous models tend to forecast the smoothed results with the averaged MAPE objectives, e.g., traffic transformer is inclined to achieve smooth predictions with hysteresis, while CD-STGNet can capture accurate sharp changes and relatively well-informed of sharp changes. This observation indicates the rationality of our core idea, anti-smooth and distinguishable prediction by exploiting context factors for capturing changes in trends.

Prediction comparisons on holiday and non-holiday. Since emerging events such as holidays can better reflect the interpretation of model outputs, we exploit our CD-STGNet to predict two series of intervals in SIP, respectively on non-holidays and **the Lantern Festival** (a traditional Chinese Festival after Chinese New Year) for comparison. The visualized prediction results are in Figure 8. Specifically, we select two consecutive intervals during 10:00-11:00 a.m., and showcase the pairwise mobility intensity maps at both non-holiday day (Jan 19th) and holiday (Feb 11th, the Lantern Festival). It is explicitly observed that the human mobility during holiday covers a larger scale than that on non-holiday, and the mobility intensity also reveals higher values when compared with intensity on non-holiday. The circles outlined with dashed and continuous lines respectively highlight observations for comparison, and such observation can exactly reflect the inherent mobility regularity in popular holidays, which enhance the understanding and interpretation of our model outputs.

4.6 Efficiency issue

Overall, our solution engages a new loss function and a novel neural structure regarding temporal evolution learner,

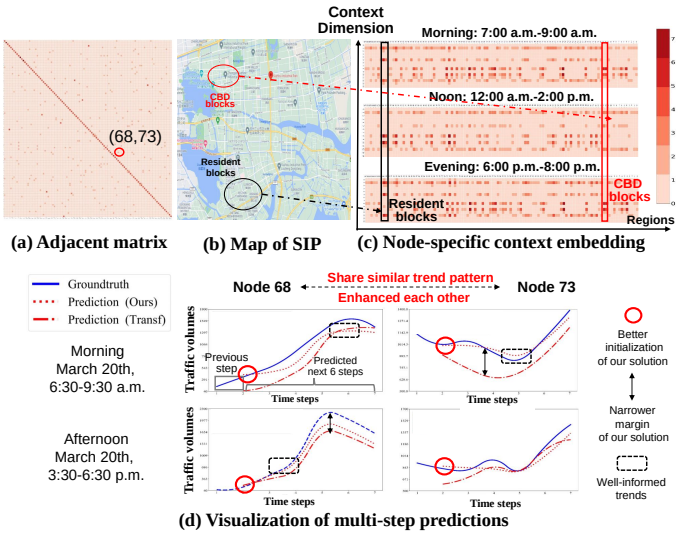


Fig. 7: Evidence of heterogeneous context response, intermediate interpretable results and final prediction comparisons. Noted that 'Transf' denotes the best baseline of Traffic Transformer on SIP.

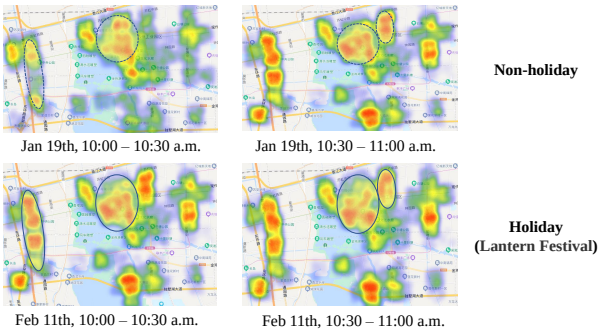


Fig. 8: Prediction comparison on holiday and non-holiday

and in fact, this solution will not introduce more computation burdens. To reveal the training costs across different models, we calculate the parameter volumes of each deep model trained on Metr-LA and present the comparison in Table 4. We can observe that our solution is with less parameters than 6 of 10 deep learning methods, thus we believe our solution can remain efficient with superior performances.

4.7 Hyperparameter study

The hyperparameters in our work are three-fold, i.e., location embedding dimensions d_{loc}, context embedding dimension K, as well as task-wise weight tuple $\gamma_i (i = 1, 2, 3, 4)$ balancing importances and magnitude orders of four optimization objectives. Note that numbers of GCN layers and LSTM layer are followed with some recent work [4]. We demonstrate the tuning process of the dimension of semantic context embedding K and location encoding d_{loc} in Table 5 and 6, respectively. Considering the above results, we choose $d_{loc} = (16, 6, 32)$ and $K = (64, 8, 32)$ respectively for SIP, NYC and Metr-LA datasets.

To analyze our adaptive optimization scheme, we report the dynamic task-wise weights during the stage of full parameter optimization in Figure 9. It is illustrated, in SIP, four tasks of MAPE, Shape-trend, Direction and Covariance, are initialized as a real-valued weight and then be updated along the learning process. The weights finally become stable at 0.51, 0.34, 0.0004 and 0.00017 after approximate 12 epochs. The weights of the direction learner and covariance are far less than others, because the direction learning task has been well trained during the first stage, and the smaller weight of covariance loss are imposed to balance each other as the graph-level covariance is usually a very large value. Instead of employing a computation costly parameter searching strategy, our adaptive optimization strategy can exactly seamlessly integrate four tasks and stabilize the learning process efficiently, resulting in tackling the intractable but ubiquitous multi-task training challenge.

TABLE 4: Parameter comparisons on different models (Unit: Million (M))

Models	Traffic Transformer	ST-SSL	STG2Seq	AGCRN	STFGNN
Parameter #	0.22M	0.22M	1.42M	0.75M	0.84M
Models	MDL	Graph-WaveNet	MTGNN	MRA-BGCN	Ours
Parameter #	0.44M	0.28M	0.41M	0.73M	0.37M

TABLE 5: Performance on various dimensions of context embedding K

K	16	32	64	96	128
SIP	0.255	0.244	**0.228**	0.255	0.289
K	4	8	16	32	64
NYC	0.135	**0.113**	0.122	0.136	0.145
Metr-LA	0.175	0.163	0.143	**0.103**	0.123

TABLE 6: Performance on various dimensions of location embedding d_{loc}

d_{loc}	8	16	32	64
SIP	0.243	**0.225**	0.238	0.265
d_{loc}	6	8	16	32
NYC	**0.113**	0.145	0.165	0.167
Metr-LA	0.126	0.121	0.114	**0.084**

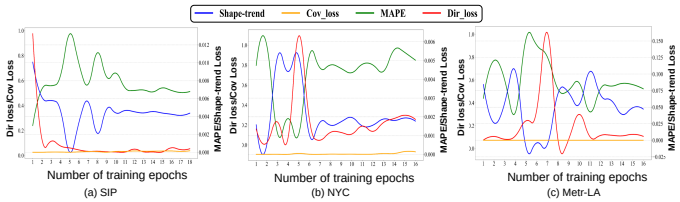


Fig. 9: Averaged task-wise loss weighting along the number of training epochs

5 Discussion

5.1 Insights of discovering human mobility regularity

Human mobility is partially predictable due to daily routines of human, however, the various context influences and

human random behavior can be root causes of spatiotemporal heterogeneity in dynamic mobility. Investigating how context factors influence human daily routines and hence mobility patterns, offers insights on designing policies to optimize the urban operations and invigorate the mobile computing in a context-aware manner. Specifically, administrations can figure out personalized scheduling or controlling strategies to avoid congestions and gathering events in different context scenarios. It is also interesting to discover the heterogeneity across different mobility types that taxicabs tend to focus on downtown while all-type vehicles are distributed more evenly and widely. Hence, it provides implications on type-specific transportation system optimization, e.g., ride-sharing system can leverage collective mobility to decide the order assignment and plan cruising routes, while administrations can exploit the regularity of type-specific mobility to formulate the vehicle-aware traffic limitations. For a general mobile computing, the Internet of vehicles can exploit the regularity of mobility to maximize the V2V transmission efficiency and balance the transmission offloading with context awareness, and the automatic pilots can optimize their routes to avoid congestions and difficult road segments in open road networks. Therefore, our findings can significantly facilitate the human-centered data mining in mobile computing community.

5.2 Technical novelty and mobile computing insights

CD-STGNet is a systematical work that not only demonstrates the necessity of context-aware prediction, but tackles spatiotemporal heterogeneity in mobility by maximally improving context utilization, contributing insights to data mining and mobile computing.

Regarding technical contributions, first, we theoretically analyze the necessity of exploiting task-related covariates to assist predictions, provides a paradigm to a series of context-sensitive learning tasks. Secondly, we summarize the discovery that the essence of spatiotemporal prediction is to explore how contexts influence historical spatial transitions to achieve future observations. Thus, CD-STGNet investigates novel neural modules and objectives to predict the initial variation direction of sequences, and exploit heterogeneous target-wise dependencies to advance the multi-step learning. Thirdly, we devise an alternate-and-adaptive parameter optimization strategy, which allows the task disentanglement and task-wise importance re-weighting for adaptively multiple task coherence.

The mobile computing insights can be delivered on two aspects. (i) Data source and utilization (application). The human mobility trajectories inherently come from mobile devices and traffics are indicators of vehicle density and volumes in road networks. Besides, vehicle volumes can also be exploited to transmit data package to enable V2V communication. Therefore, our proposal can be deemed as tackling the application challenge in mobile computing, i.e., how to effectively exploit the data collected from mobile devices. (ii) Technical insights. The main characteristic of mobile computing data is dynamic and heterogeneous, correspondingly, the context-directional data aggregation mechanism in CD-STGNet opportunely models these two features in mobile data. Such learning-based prediction can facilitate the development of mobile computing techniques in communications, e.g., modeling the heterogeneity to balance the task/computation offloading. Therefore, we can provide new technologies for the communication optimization at bottom layer of mobile computing, to enable efficient and predictable communication. To summarize, we believe CD-STGNet can benefit the broad audience of mobile computing community.

5.3 Limitations

CD-STGNet has answered the question of which direction should each node propagate, as well as how to exploit multiple contexts and their interactions to perform personalized spatial-and-temporal aggregations. Such solution will gain more bonus when the temporal fluctuation is larger. To this end, CD-STGNet doesn't compete against MTGNN on Metr-LA, which is mostly because that the data fluctuation of Metr-LA tends to be more moderate than other two datasets. Therefore, the first underlying weakness of our solution is that the optimization process of learnable parameters must depend on the temporal evolution where lower temporal dynamics can lead to similar node selection pattern and indistinguishable aggregation. Secondly, this solution still cannot quantify the node propagation steps on spatial learning, and cannot avoid the inherent cascaded errors between spatial aggregations and temporal learning. For this, the promising solution is to realize a dynamic network to control the propagation depth by exploiting auxiliary information, and a collaboration strategy with information sharing can be further studied for cascaded error alleviation.

6 Related Work

6.1 Human Mobility Prediction

Human mobility prediction. Human beings are routine-oriented, which enables high predictability in daily mobility patterns [40]. Research of human mobility predictions can be categorized into two levels, individual-level trajectories and collective mobility. Individual-level prediction aims to forecast the user-specific mobility and activity by investigating their historical records. Specifically, traditional Markov Chain [41], variants of RNN, e.g., GRU [42] and LSTM [43] are modified to capture individual transition regularity. Nevertheless, these works cannot identify the overall status of POIs for avoiding collective urban events, and these solutions consider all users equally, consequently failing to explicitly take user personalization and spatiotemporal heterogeneity in predictions. Collective human mobility is highly concerning with predictions of traffic volumes [1], [12], [24], taxicab trips [7], [22] and collective urban events [3], [31]. These tasks have been resolved with advanced CNN [1], [22] or GNN [12], [14], [15]. In particular, [3] devises a model-ensemble solution to address the spatial heterogeneity in state-wide risk predictions, while [31] introduces a dynamic adjacent matrix and difference operator to perceive the time-varying changes of collective mobility. Even though, the context-induced heterogeneity of human mobility has been less explored, regardless of collective or individual ones. Promisingly, as environmental context factors can significantly influence human daily routines [44], [45], the uncertainty in individual-level predictions inspires us to leverage

context factors for further improving the predictability of collective mobility.

6.2 Spatiotemporal Forecasting

Collective mobility prediction is a typical spatiotemporal learning task. Existing techniques of spatiotemporal learning can be classified into context-agnostic and context-involved solutions. Within the former class, literature [16], [46] introduce multi-view graphs to encode cross-view spatial correlations, while [12] and [47] design an attention-based spatiotemporal encoder to capture the spatial heterogeneity. Unfortunately, both of them ignore the critical roles of informative contexts in forecasting and yield suboptimal performance. With the increasing awareness of contexts, recent context-involved solutions employ a two-layer fully connected neural network to encode the contextual information and then aggregate them with main features, where the fusion mechanisms can be particularly classified into element-wise addition [1], [15], [16], [39], [48] and vector-wise concatenation [22], [49]. More recently, ST-SSL injects the geographical contexts by performing convolution on region embedding [36]. However, given the context-induced spatiotemporal heterogeneity, there are two critical issues neglected in above works: 1) They still have not considered the interactions of context factors, which can be root causes of various heterogeneity. 2) They fail to exploit the guidance of context on spatiotemporal aggregation, leading to homogeneous aggregation strategies, and suboptimal performance. To summarize, exploring appropriate solutions to capturing context-wise interactions and influences is essential for spatiotemporal prediction.

6.3 Target Dependence Learning

Machine learning researchers start to exploit label correlations to improve learner performances. Gen, et, al. [50] firstly propose the label distribution learning, to reconstruct distribution of labels, enabling the network to alleviate label ambiguity. This series of works arise the attention on considering potential label distributions and explicit label correlations. To introduce label correlations into GNNs, Label-Aware GNN has been proposed to identify node-wise label correlations and filter all negative neighbors for aggregations by an edge classifier [51]. Recently, Google, which develops the Neural Process [52] and Conditional Neural Process (CNP) [26], further demonstrates that CNP can improve learning efficiency by enjoying the benefits of both prior knowledge sampling in Gaussian process and gradient-based optimization in neural networks. Excitingly, some pioneering work has theoretically demonstrated that explicitly modeling target-wise correlations can imitate CNP by minimizing the discrepancy of target variable-wise correlations between prediction results and groundtruth [25]. Nevertheless, target variable-wise correlations have never been explicitly considered in spatiotemporal forecasting, which implies an opportunity of performance improvement with modeling target-wise dependencies.

7 Conclusion

In this work, we devote to a systematical study on improving collective mobility prediction via countering spatiotemporal heterogeneity. We perform a mutual corroboration on the intuition that predictive power can be gained by incorporating context factors into learning algorithms, with both theoretical analysis and data-driven case visualizations. To tackle such spatiotemporal heterogeneity, we resort to context factor modeling and target-wise heterogeneous dependence constraining. Instead of limited neural architecture designs, we propose our CD-STGNet in the design perspectives of in-depth data utilization and innovative objectives where three contributions have been made. First, to perform node-specific aggregations, we additionally leverage the widely available context factors to realize a direction learner, generating node-wise vector fields for target-oriented neighbor selection. Second, for temporal learning, we bridge the gap between spatial feature maps and targeted sequences by disentangling three learnable transformations. In particular, two novel objectives considering element-wise shape-trend and pairwise covariances are devised to regularize and constrain the directional trends and spatial correlations consistent with groundtruth. Thirdly, the effectiveness and interpretability of CD-STGNet have been carefully verified with various experimental designs on three different types of mobility datasets. Finally, we demonstrate the interpretation of our CD-STGNet with multiple case studies and discuss the key insights of technologies and utilization of our work on mobile computing community.

For future research, we plan to develop mobility-driven mobile computing systems, such as intelligent transportation system and mobility-aware data transmission scheme to facilitate the application of advanced machine learning algorithms.

Acknowledgments

This paper is partially supported by National Natural Science Foundation of China (No.62072427,12227901,62271452), the Project of Stable Support for Youth Team in Basic Research Field, CAS (No.YSBR-005), the National Key Research and Development Program of China (No.2022YFB4500300), and Key Research Project of Zhejiang Lab (No.2022PI0AC01).

References

[1] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[2] Q. Yuan, Z. Liu, J. Li, J. Zhang, and F. Yang, "A traffic congestion detection and information dissemination scheme for urban expressways using vehicular networks," *Transportation Research Part C: Emerging Technologies*, vol. 47, pp. 114–127, 2014.

[3] Z. Yuan, X. Zhou, and T. Yang, "Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 984–992.

[4] Z. Zhou, Y. Wang, X. Xie, L. Chen, and C. Zhu, "Foresee urban sparse traffic accidents: A spatiotemporal multi-granularity perspective," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[5] Y. Wang, Z. Lv, W. Chen, and H. Liu, "Wi-eye: Tracking urban private vehicles with inter-vehicle communications and sparse video surveillance cameras," in *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2018, pp. 1–9.

[6] Y. Wang, E. Yang, W. Zheng, L. Huang, H. Liu, and B. Liang, "A realistic and optimized v2v communication system for taxicabs," in *2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2016, pp. 139–148.

[7] Y. Zhang, B. Wang, Z. Shan, Z. Zhou, and Y. Wang, "Cmt-net: A mutual transition aware framework for taxicab pick-ups and drop-offs co-prediction," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022, pp. 1406–1414.

[8] Y. Wang, Z. Zhou, K. Liu, X. Xie, and W. Li, "Large-scale intelligent taxicab scheduling: A distributed and future-aware approach," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 8176–8191, 2020.

[9] W. Chen, L. Chen, Y. Xie, W. Cao, Y. Gao, and X. Feng, "Multi-range attentive bicomponent graph convolutional network for traffic forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3529–3536.

[10] X. Li, R. Hu, Z. Wang, and T. Yamasaki, "Location prediction via bi-direction speculation and dual-level association," *IJCAI 2021*, 2021.

[11] H. Ren, Y. Song, J. Wang, Y. Hu, and J. Lei, "A deep learning approach to the citywide traffic accident risk prediction," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3346–3351.

[12] S. Guo, Y. Lin, N. Feng, C. Song, and n. H. Wa, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 922–929.

[13] Y. Liang, K. Ouyang, J. Sun, Y. Wang, J. Zhang, Y. Zheng, D. Rosenblum, and R. Zimmermann, "Fine-grained urban flow prediction," in *Proceedings of the Web Conference 2021*, 2021, pp. 1833–1845.

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

[15] L. Bai, L. Yao, S. S. Kanhere, X. Wang, and Q. Z. Sheng, "Stg2seq: spatial-temporal graph to sequence model for multi-step passenger demand forecasting," in *28th International Joint Conference on Artificial Intelligence, IJCAI 2019*. International Joint Conferences on Artificial Intelligence, 2019, pp. 1981–1987.

[16] M. Li and Z. Zhu, "Spatial-temporal fusion graph neural networks for traffic flow forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 5, 2021, pp. 4189–4196.

[17] Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, and C. 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*, 2020, pp. 753–763.

[18] S. Wang, H. Miao, H. Chen, and Z. Huang, "Multi-task adversarial spatial-temporal networks for crowd flow prediction," in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 1555–1564.

[19] W. Jiang and J. Luo, "Graph neural network for traffic forecasting: A survey," *Expert Systems with Applications*, p. 117921, 2022.

[20] J. Ji, J. Wang, Z. Jiang, J. Jiang, and H. Zhang, "Stden: Towards physics-guided neural networks for traffic flow prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 4, 2022, pp. 4048–4056.

[21] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[22] J. Ye, L. Sun, B. Du, Y. Fu, X. Tong, and H. Xiong, "Co-prediction of multiple transportation demands based on deep spatio-temporal neural network," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 305–313.

[23] L. Wang, D. Chai, X. Liu, L. Chen, and K. Chen, "Exploring the generalizability of spatio-temporal traffic prediction: meta-modeling and an analytic framework," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[24] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[25] B. Yoo, J. Lee, J. Ju, S. Chung, S. Kim, and J. Choi, "Conditional temporal neural processes with covariance loss," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 051–12 061.

[26] M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. Rezende, and S. A. Eslami, "Conditional neural processes," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1704–1713.

[27] J.-H. Zhai, L.-Y. Wan, and M.-Y. Zhai, "Attribute reduction based on the principle of maximal dependency and minimal mutual information," in *2012 International Conference on Machine Learning and Cybernetics*, vol. 1. IEEE, 2012, pp. 272–276.

[28] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *European semantic web conference*. Springer, 2018, pp. 593–607.

[29] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[30] D. Beaini, S. Passaro, V. Létourneau, W. Hamilton, G. Corso, and P. Liò, "Directional graph networks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 748–758.

[31] Z. Zhou, Y. Wang, X. Xie, L. Chen, and H. Liu, "Riskoracle: A minute-level citywide traffic accident forecasting framework," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 1258–1265.

[32] Z. Pan, Y. Liang, W. Wang, Y. Yu, Y. Zheng, and J. Zhang, "Urban traffic prediction from spatio-temporal data using deep meta learning," in *Proceedings of the 25th ACM SIGKDD*, 2019, pp. 1720–1730.

[33] P. Tseng and S. Yun, "A coordinate gradient descent method for nonsmooth separable minimization," *Mathematical Programming*, vol. 117, no. 1, pp. 387–423, 2009.

[34] J. Kingma, D.P.; Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] L. Cai, K. Janowicz, G. Mai, B. Yan, and R. Zhu, "Traffic transformer: Capturing the continuity and periodicity of time series for traffic forecasting," *Transactions in GIS*, vol. 24, no. 3, pp. 736–755, 2020.

[36] J. Ji, J. Wang, C. Huang, J. Wu, B. Xu, Z. Wu, J. Zhang, and Y. Zheng, "Spatio-temporal self-supervised learning for traffic flow prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[37] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph wavenet for deep spatial-temporal graph modeling," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 1907–1913.

[38] S. Abu-El-Haija, B. Perozzi, A. Kapoor, N. Alipourfard, K. Lerman, H. Harutyunyan, G. Ver Steeg, and A. Galstyan, "Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing," in *international conference on machine learning*. PMLR, 2019, pp. 21–29.

[39] J. Zhang, Y. Zheng, J. Sun, and D. Qi, "Flow prediction in spatio-temporal networks based on multitask deep learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 3, pp. 468–478, 2019.

[40] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific reports*, vol. 3, p. 2923, 2013.

[41] R. He and J. McAuley, "Fusing similarity models with markov chains for sparse sequential recommendation," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 2016, pp. 191–200.

[42] C. Yang, M. Sun, W. X. Zhao, Z. Liu, and E. Y. Chang, "A neural network approach to jointly modeling social networks and mobile trajectories," *ACM Transactions on Information Systems (TOIS)*, vol. 35, no. 4, pp. 1–28, 2017.

[43] P. Zhao, A. Luo, Y. Liu, F. Zhuang, J. Xu, Z. Li, V. S. Sheng, and X. Zhou, "Where to go next: A spatio-temporal gated network for next poi recommendation," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[44] A. Cuttone, S. Lehmann, and M. C. González, "Understanding predictability and exploration in human mobility," *EPJ Data Science*, vol. 7, pp. 1–17, 2018.

[45] M. De Domenico, A. Lima, and M. Musolesi, "Interdependence and predictability of human mobility and social interactions," *Pervasive and Mobile Computing*, vol. 9, no. 6, pp. 798–807, 2013.

[46] X. Geng, Y. Li, L. Wang, L. Zhang, Q. Yang, J. Ye, and Y. Liu, "Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3656–3663.

[47] S. Guo, Y. Lin, H. Wan, X. Li, and G. Cong, "Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting," *IEEE Transactions on Knowledge and Data Engineering*, 2021.
[48] L. Bai, L. Yao, X. Wang, C. Li, and X. Zhang, "Deep spatial–temporal sequence modeling for multi-step passenger demand prediction," *Future Generation Computer Systems*, vol. 121, pp. 25–34, 2021.
[49] J. Bao, P. Liu, and S. V. Ukkusuri, "A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data," *Accident Analysis & Prevention*, vol. 122, pp. 239–254, 2019.
[50] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.
[51] H. Chen, Y. Xu, F. Huang, Z. Deng, W. Huang, S. Wang, P. He, and Z. Li, "Label-aware graph convolutional networks," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1977–1980.
[52] M. Garnelo, J. Schwarz, D. Rosenbaum, F. Viola, D. J. Rezende, S. Eslami, and Y. W. Teh, "Neural processes," *arXiv preprint arXiv:1807.01622*, 2018.

Zhengyang Zhou is now an associate researcher at Suzhou Institute for Advanced Research, University of Science and Technology of China (USTC). He got his Ph.D. degree at University of Science and Technology of China in 2023. He has published over 20 papers on top conferences and journals such as KDD, ICLR, TKDE, WWW, AAAI, SDM and ICDE. His mainly research interests include human-centered urban computing, and mobile data mining.

Kuo Yang received his BE degree from Northeastern University at Qinhuangdao, China, in 2021. He is now a PhD candidate at School of Data Science, USTC. His mainly research interests are data-driven urban management, spatiotemporal data mining and mobile computing.

Yuxuan Liang is an Assistant Professor at Intelligent Transportation Thrust, Hong Kong University of Science and Technology (Guangzhou). He is currently working on the research, development, and innovation of spatio-temporal data mining and AI, with a broad range of applications in smart cities. Prior to that, he obtained his PhD degree at NUS. He published over 40 peer-reviewed papers in refereed journals and conferences, such as KDD, WWW, NeurIPS, ICLR, ECCV, AAAI, IJCAI, Ubicomp, and TKDE. Those papers have been cited over 2,100 times (Google Scholar H-Index: 21). He was recognized as 1 out of 10 most innovative and impactful PhD students focusing on data science in Singapore by Singapore Data Science Consortium (SDSC).

Binwu Wang is now a PhD candidate at School of Data Science, USTC. His research interests include spatial-temporal data mining and human-centered urban computing. He has published over 5 refereed journal and conference papers in the field of data mining, including IEEE TITS, DASFAA, IEEE ICDM, WSDM, etc.

Hongyang Chen received his B.S. and M.S. degrees from Southwest Jiaotong University, China, in 2003 and 2006, and Ph.D. degree from University of Tokyo, Japan, in 2011. He is currently a Senior Research Expert with Zhejiang Lab, China. He has authored 100+ refereed journal and conference papers in ACM Sensor Networks, IEEE TSP, IEEE TWC and IEEE ICC. His research interests include IoT, data-driven intelligent systems, machine learning, location-based big data, and statistical signal processing.

Yang Wang is now an associate professor at USTC. He got his Ph.D. degree at University of Science and Technology of China in 2007. He has published over 100 high-level conference and journal papers on IEEE TKDE, ICLR, IJCAI, AAAI, MOBICOM, WWW, et, al. His research interest mainly includes distributed system, urban computing, spatiotemporal data mining, and data-driven interdisciplinary research. He is also a senior member of both ACM and IEEE.