Learning from Highly Sparse Spatio-temporal Data

Leyan Deng[†], **Defu Lian** [†§*], **Chenwang Wu**[†], **Enhong Chen**[†§]
[†] School of Artificial Intelligence and Data Science, University of Science and Technology of China
[§] School of Computer Science and Technology, University of Science and Technology of China
{dleyan, wcw1996}@mail.ustc.edu.cn, {liandefu, cheneh}@ustc.edu.cn

Abstract

Incomplete spatio-temporal data in the real world has spawned much research. However, existing methods often utilize iterative message-passing across temporal and spatial dimensions, resulting in substantial information loss and high computational cost. We provide a theoretical analysis revealing that such iterative models are susceptible to data and graph sparsity, causing unstable performances on different datasets. To overcome these limitations, we introduce a novel method named One-step Propagation and Confidence-based Refinement (OPCR). In the first stage, OPCR leverages inherent spatial and temporal relationships by employing a sparse attention mechanism. These modules propagate limited observations directly to the global context through one-step imputation, which is theoretically affected only by data sparsity. Following this, we assign confidence levels to the initial imputations by correlating missing data with valid data. This confidence-based propagation refines the separate spatial and temporal imputation results through spatio-temporal dependencies. We evaluate the proposed model across various downstream tasks involving highly sparse spatio-temporal data. Empirical results indicate that our model outperforms state-of-the-art imputation methods, demonstrating its effectiveness and robustness.

1 Introduction

Spatio-temporal data, encompassing information distributed across both spatial and temporal dimensions, is prevalent in various domains [1]. In real world, the acquisition of spatio-temporal data often faces practical constraints, leading to missing observations. Incomplete data undermines the reliability and effectiveness of subsequent tasks, necessitating robust imputation techniques. However, the existing spatio-temporal works effort addresses temporal missing (i.e., point-level missing), which is usually caused by inevitable device failure. Beyond that, spatial missing (i.e., node-level missing) becomes a burning issue. For example, in a traffic scenario, the management will not deploy dense traffic sensors to reduce costs. Therefore, exploring spatially sparse data not only makes the best use of available data but also lowers the barriers to implementing spatio-temporal models in real scenarios. In view of the gap, Traffic4cast 2022 (T4C22) competition [2] presented sparse traffic data collected in real urban road networks. Specifically, the competition considered three cities, where sensors are deployed sparsely, i.e. spatial and point missing co-exist. T4C22 aims at generalizing limited data to entire city to predict global traffic dynamics.

We first provide a theoretical analysis for general spatio-temporal iterative imputation model from PAC-learnability perspective. Then we find that iterative methods are constrained by the sparsity and structure of data. As the data scale and missing rate increase, the model requires more iterations. Additionally, with an increasing number of iterations, the model requires more training samples. Therefore, for highly sparse large-scale data, iterative methods suffer from information loss and error accumulation. To address these issues, we propose a sparse attention-based one-step imputation and confidence-based refinement approach. In the first stage, we propagate information from observed

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2982

data to all missing data directly, leveraging inherent spatial and temporal correlations. In the second stage, we assign confidence-based propagation weights to the imputed results. Through confidence-based refinement, we eliminate the bias introduced by imputations from separate perspectives.

This paper makes the following contributions.

- We provide an limitation analysis for the popular iterative imputation from the PAC-learnability perspective, which explains the error accumulation caused by multiple iterations.
- Motivated by theoretical results, we propose a one-step propagation strategy to efficiently recover global information from limited observations. Then we assign confidence-based propagation weights to spatio-temporal interactions, refining the imputation results.
- We apply our method to various downstream tasks, the experiments show that our model can learn sufficient information from limited observations than state-of-the-art baselines.

2 Related Works

There are many works exploring how to impute different types of sparse data. For tabular data, matrix factorization-based methods are widely popular. A recent work transformed tabular data imputation into link prediction in a bipartite graph, which benefited from the expressive power of graph neural networks (GNNs). For graph data, a simple idea is propagating known features to missing data through the graph structure. For example Feature propagation (FP) [3] iteratively aggregates neighboring node features and PCFI [4] further considered pseudo-confidence to propagate messages more accurately. It is clear that GNN can capture spatial relations better. In addition to GNN-based imputation model, SAT [5] and ITR [6] both additionally considered structural representations and structure reconstruction objective. However, these graph imputation problems focus on static graph and fixed missing data sets, which is incompatible with dynamic sparse spatio-temporal data.

For time-series data, the existing approaches investigated the imputation performances of different sequence models. For example, BRITS [7] proposed a bidirectional Recurrent Neural Networks (RNN). SAITS [8] designed self-attention based model and jointly trained imputation and reconstruction objective. CSDI [9] proposed a conditional score-based diffusion model. However, these time-series imputation methods ignore intrinsic spatial dependencies in spatio-temporal data.

For spatio-temporal data, existing research leverages topological information effectively. IGNNK [10] introduced a scalable inductive method trained by reconstructing random sub-graphs. GRIN [11] was the first to apply Graph Neural Networks (GNNs) to multivariate time-series imputation and proposed a graph-based recurrent neural network. Subsequently, SPIN [12] identified that auto-regressive models like GRIN are prone to error propagation. SPIN developed an end-to-end architecture utilizing intra-node and inter-node attention mechanisms to address this. However, when dealing with highly sparse large-scale data, propagating valid information through graph structures requires multiple iterations, leading to error accumulation and information loss. In addition, PriSTI [13] proposed a diffusion-based model. Nonetheless, PriSTI employs two separate attention modules to incrementally aggregate temporal and spatial dependencies, which leads to decoupling the spatio-temporal context. Additionally, PriSTI uses linear interpolation for coarse conditional information, which is inadequate for spatially missing data. Beyond attribute imputation, PoGeVon [14] also addresses the challenge of missing structural information in spatio-temporal data.

3 Preliminaries

Definition 3.1 (**Spatio-temporal Series**)**.** Given a fixed graph $G = (\mathcal{V}, \mathcal{E}, \boldsymbol{A})$, where $\mathcal{V}$ and $\mathcal{E}$ denote the set of N nodes and E edges, respectively; $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ is the corresponding adjacency matrix. A spatio-temporal series of length T is represented by $\boldsymbol{X} \in \mathbb{R}^{N \times T \times F_x}$, where F_x denotes the number of feature dimensions. We represent each spatio-temporal point (ST point) with a tuple $z = (v, t)$ and $\boldsymbol{x}_{v,t} \in \mathbb{R}^{F_x}$ denotes the collected data of node v at time t.

Definition 3.2 (**Spatio-temporal Mask**)**.** For spatio-temporal series $\boldsymbol{X}$, we use a binary mask $\boldsymbol{M}$ to indicate missing ST points, where $m_{v,t} = 0$ if $\boldsymbol{x}_{v,t}$ is missing, otherwise $m_{v,t} = 1$. Note that, since a faulty device often miss all records, we ignore the availability of data in feature dimension. Thus,

we can define a unique sparse spatio-temporal series $\tilde{\boldsymbol{X}}$ in terms of $\boldsymbol{X}$ and $\boldsymbol{M}$, i.e.,

$$\tilde{\boldsymbol{X}} = \boldsymbol{X} \odot \boldsymbol{M} + NaN \odot (\boldsymbol{1} - \boldsymbol{M}), \tag{1}$$

Definition 3.3 (**Sparse Spatio-temporal Data Learning**). The goal is to accomplish downstream tasks based on the incomplete data $\tilde{\boldsymbol{X}}$. Without loss of generality, we consider the following objective functions for node-level task.

$$\mathcal{L}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \frac{1}{N} \sum_{v \in \mathcal{V}} l(\boldsymbol{y}_v, \hat{\boldsymbol{y}}_v), \tag{2}$$

where $l(\cdot, \cdot)$ is an element-wise loss function that depends on specific downstream tasks. Here $\boldsymbol{y}_v$ and $\hat{\boldsymbol{y}}_v$ are the ground truth and model's prediction for node v, respectively.

For imputation task, the model aims to minimize the reconstruction error as follows.

$$\mathcal{L}(\tilde{\boldsymbol{X}}, \hat{\boldsymbol{X}}) = \frac{1}{|\boldsymbol{1} - \boldsymbol{M}|} \sum_{v \in \mathcal{V}} \sum_{t \in \mathcal{T}} (1 - m_{v,t}) \cdot l(\boldsymbol{x}_{v,t}, \hat{\boldsymbol{x}}_{v,t}), \tag{3}$$

where $\mathcal{T}$ denotes the set of time steps and $l(\cdot, \cdot)$ is an element-wise loss function.

4 Theoretical Analysis

As we stated above, the iterative spatio-temporal imputation methods will suffer from information loss and error accumulation on extremely sparse large-scale data. In order to expose the limitations of iterative models, we provide a theoretical analysis in terms of PAC-learnability for node-level tasks.

We use $\boldsymbol{X} \in \mathcal{X}$ and $\boldsymbol{Y} \in \mathcal{Y}$ to denote complete spatio-temporal series and node-level labels, and use $\boldsymbol{M} \in \mathcal{M}$ to denote mask distribution. Let $\mathcal{D}$ be the fixed but unknown distribution over $\mathcal{X} \times \mathcal{Y}$. We define the mixture distribution $\tilde{\mathcal{D}}$ of $\mathcal{D}$ and $\mathcal{M}$ using Eq. 1. Given a training set $\tilde{S} = \{(\tilde{\boldsymbol{X}}_i, \boldsymbol{Y}_i)\}^m$, we assume that all samples in $\tilde{S}$ are i.i.d. according to $\tilde{\mathcal{D}}$, denoted as $\tilde{S} \sim \tilde{\mathcal{D}}^m$. Let $f \in \mathcal{F} : \mathcal{X} \times \mathcal{M} \to \mathcal{Y}$ be a sparse spatio-temporal data learning model, the empirical risk over $\tilde{S}$ and the corresponding generalization risk is defined as follows,

$$\textbf{Empirical risk} : \hat{\mathcal{R}}_{\tilde{S}}(f) = \frac{1}{m} \sum_{i=1}^{m} l(f(\tilde{\boldsymbol{X}}_i), \boldsymbol{Y}_i),$$

$$\textbf{Generalization risk} : \mathcal{R}_{\tilde{D}}(f) = \mathbb{E}_{(\tilde{\boldsymbol{X}}, \boldsymbol{Y}) \sim \tilde{D}} \left[l(f(\tilde{\boldsymbol{X}}), \boldsymbol{Y}) \right]$$

Similarly, we denote empirical and generalization risk under complete data as $\hat{\mathcal{R}}_S(f)$ and $\mathcal{R}_{\mathcal{D}}(f)$.

Definition 4.1 (**PAC-Learnability**[15]). A concept class $\mathcal{C}$ is said to be probably approximately correct (PAC) learnable if there exist a learning algorithm $\mathcal{A}$ and a polynomial function $poly(\cdot, \cdot, \cdot, \cdot)$ satisfying: for any $\epsilon, \delta > 0$ and any distribution $\mathcal{D}$, the following holds for any sample size $m \geq poly(1/\epsilon, 1/\delta, n, size(c))$:

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left[R_{\mathcal{D}}(h_S) \leq \epsilon \right] \geq 1 - \delta.$$

Most of sparse spatio-temporal data learning models have the following encoder-decoder architecture. The encoder f_e learns all ST point representations from sparse data and the decoder f_d outputs node-level predictions for specific downstream task.

$$f_d(f_e(\tilde{\boldsymbol{X}})) = \phi \left(\frac{1}{T} \sum_{t \in \mathcal{T}} f_e(\tilde{\boldsymbol{X}})_t \boldsymbol{W}_d \right),$$

where $\boldsymbol{W}_d$ represents the trainable weight matrix in decoder. In the encoder, the iterative models to learn the missing ST point representations work by continuously extracting valid information from its neighbors. The key differences between iterative models lie in how they define neighbors and how they aggregate messages from these neighbors.

Let $\mathcal{Z} = \{(v, t) | v \in \mathcal{V}, t \in \mathcal{T}\}$. For any ST point $z \in \mathcal{Z}$, we denote the set of its neighboring ST points at the k-th iteration by $\mathcal{N}_z^k$. According to the definition of neighbors, we can infer the path

https://doi.org/10.52202/079017-2982

between any two ST points. Without loss of generality, for any $z \in \mathcal{Z}$, we can formulate the k-th iteration as follows.

$$\boldsymbol{h}_z^k = \phi\left(\sum_{z' \in \mathcal{N}_z^k} p_{z' \to z}^k \cdot \boldsymbol{h}_{z'}^{k-1}\right),$$

where $\boldsymbol{h}_z^k$ denotes the learned representation for ST point z after the k-th iteration. For message-passing, the iterative models usually assign weight $p_{z' \to z}^k$ to messages from z' to z. We denote the distance of the path from ST point z to z' as $d_{z \to z'}$. Let $\mathcal{Z}_o \subseteq \mathcal{Z}$ to be the set of all observed ST points and $\mathcal{Z}_m = \mathcal{Z} \backslash \mathcal{Z}_o$. To ensure that the model has ability to recover all ST points, the number of iterations must satisfies $K \geq \max_{z' \in \mathcal{Z}_m} \min_{z \in \mathcal{Z}_o} d_{z \to z'}$. Therefore, multiple iterations are inevitable in highly sparse large-scale spatio-temporal data.

We then establish the PAC-learnable guarantee for the iterative spatio-temporal imputation methods. First, let us immediately mask some mild assumptions that are easy to implement.

Assumption 4.2. For the weight matrix in decoder, we assume its spectral norm satisfies $\|\boldsymbol{W}_d\|_2 \leq B_d$, for any ST point $z \in \mathcal{Z}$, we assume its feature vector satisfies $\|\boldsymbol{x}_z\|_2 \leq B_x$. For the loss function, we assume l is an C_l-lipschitz continuous and bounded by $[0, 1]$. For the activation function, we assume ϕ is C_ϕ-lipschitz continuous with $\phi(0) = 0$. For the mask distribution $\mathcal{M}$, we assume that all ST points are randomly masked with a probability of ρ.

Proposition 4.3. *Let $\mathcal{F}$ be a K-iterations imputation model class, we assume its learned propagation weights are bounded by $[0, \gamma]$. If we draw a sample $\tilde{S}$ of size m, for any $f \in \mathcal{F}$, the following inequality holds.*

$$\mathbb{P}_{\tilde{S} \sim \tilde{\mathcal{D}}^m}\left[R_{\tilde{\mathcal{D}}}(h_{\tilde{S}}) \leq \epsilon\right] \geq 1 - \delta$$

with $m \geq \frac{\log 1/\delta}{2\epsilon - 4\eta\mathcal{C}K - 4(1-\eta)\mathcal{C}K \cdot \rho^\tau - 4C_l\Re_m(\mathcal{F})}$, where $\mathcal{C} = C_l \cdot \gamma^K \cdot C_\phi^{K+1} \cdot (d_{\max})^K \cdot B_x B_d$.

Remark 4.4. Here $0 < \eta < 1$, $d_{\max} = \max_{z \in \mathcal{Z}} |\mathcal{N}_z|$ and $\tau = \max_{z \in \mathcal{Z}} |\{d_{z' \to z} < K | z' \in \mathcal{Z}\}|$. Consider PAC-learnability under complete data, number of required samples needs to satisfy $m \geq \frac{\log 1/\delta}{2(\epsilon - 2C_l\Re_m(\mathcal{F}))}$. Since $\Re_m(\mathcal{F})$ represents Rademacher complexity of model class $\mathcal{F}$, it is encouraging that the learnability of the model on sparse data is related to the one on complete data. Thus, we can estimate the impact of the missing data on the model performance. Obviously, the required sample size m is positively correlated with the missing ratio ρ and the number of model iterations K. Nevertheless, the structural sparsity in the spatio-temporal data itself limits the learnability of the iterative model in addition to the data sparsity. Specifically, τ is related to the number of k-hop neighbors of each ST point. Therefore, the sparser the spatio-temporal structure (i.e., smaller τ), the greater the number of samples required. This theoretical result provides an interpretation for iterative model-induced error accumulation. Note that although we assume that all ST points are masked at random (MAR), this proof strategy can be generalized to other mask distribution $\mathcal{M}$.

5 Methodology

This section presents our proposed method, One-step Propagation and Confidence-based Refinement (OPCR), as illustrated in Fig. 1. In the first stage, we independently learn intrinsic representations for ST points in both spatial and temporal dimensions. We then employ a sparse attention-based one-step propagation strategy to obtain two separate imputation results. In the second stage, we derive imputation confidence based on the learned correlations between ST points to address potential biases in these results. Finally, we perform confidence-based spatio-temporal propagation to refine the final predictions.

5.1 Sparse Attention-based One-step Propagation

Motivated by the aforementioned theoretical findings, we propose a one-step propagation method to avoid error accumulation and information loss commonly found in iterative methods. Specifically, as illustrated in Fig. 1, we begin by independently learning intrinsic spatial and temporal representations for ST points. We capture correlations among ST points across different dimensions by adopting a sparse attention mechanism. This approach enables the limited information to be propagated directly to the global context as efficiently as possible.

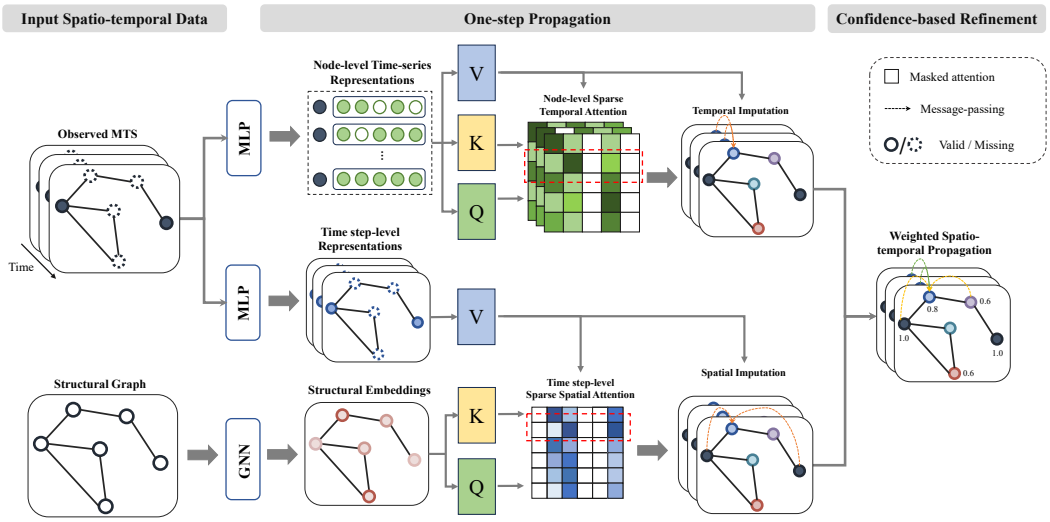

Figure 1: The framework of the proposed OPCR.

5.1.1 Sparse Spatial Attention

We first address the significant spatial sparsity present in spatio-temporal data. To fully utilize limited observation, we learn the correlations between nodes using static spatial information. Leveraging these intrinsic dependencies, messages from all observations are propagated directly to the global context through a sparse attention mechanism. We utilize a L_s-layers Graph Neural Network (GNN) [16] to learn the spatial representations of nodes from the graph structure. We initialize $\boldsymbol{S}_0$ using the inherent static attributes of all nodes; then the layer-wise updation can be summarized as follows.

$$\boldsymbol{S}_l = \phi\left(\boldsymbol{S}_{l-1}\boldsymbol{W}_1^l + \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{S}_{l-1}\boldsymbol{W}_2^l\right),$$

where $\boldsymbol{W}_1^l$ and $\boldsymbol{W}_2^l$ are the parameter matrix in the l-layer; $\boldsymbol{D}$ is the degree matrix of $\boldsymbol{A}$.

In order learn the spatial associations between nodes, we employ a self-attention mechanism [17]. We use node embeddings $\boldsymbol{S} = \boldsymbol{S}_{L_s}$ as both the key and query. We denote the set of available nodes by $\tilde{\mathcal{V}} \subset \mathcal{V}$. For any node $u \in \mathcal{V}, v \in \tilde{\mathcal{V}}$, the spatial self-attention is formulated as:

$$\begin{aligned} \boldsymbol{Q}^s, \boldsymbol{K}^s, \boldsymbol{V}_t^s &= \boldsymbol{S}\boldsymbol{W}_Q^s, \boldsymbol{S}\boldsymbol{W}_K^s, \boldsymbol{X}_t\boldsymbol{W}_V^s, \\ \alpha_{u,v}^s &= \frac{\exp(\langle \boldsymbol{q}_u^s, \boldsymbol{k}_v^s \rangle)}{\sum_{v' \in \tilde{\mathcal{V}}} \exp(\langle \boldsymbol{q}_u^s, \boldsymbol{k}_{v'}^s \rangle)}, \end{aligned} \tag{4}$$

where $\boldsymbol{Q}^s, \boldsymbol{K}^s, \boldsymbol{V}_t^s \in \mathbb{R}^{N \times F}$ represent the query, key, value of spatial self-attention; $\boldsymbol{q}_i^s$, $\boldsymbol{k}_i^s$ and $\boldsymbol{v}_{i,t}^s$ stand for their i-the row; $\boldsymbol{W}_Q^s, \boldsymbol{W}_K^s, \boldsymbol{W}_V^s \in \mathbb{R}^{F \times F}$ are parameter matrices. For any ST point (v, t), we aggregate spatial messages coming from all observations weighted by learned attentions. The spatial dependencies-based imputation result can be computed as:

$$\boldsymbol{h}_{v,t}^s = \text{MLP}\left(\sum\nolimits_{v' \in \tilde{\mathcal{V}}} \alpha_{v,v'}^s \boldsymbol{v}_{v',t}^s\right). \tag{5}$$

5.1.2 Sparse Temporal Attention

Another component is designed to learn temporal correlations. We adopt a temporal attention mechanism for the time-series data associated with each node. Unlike recurrent sequence models such as RNN or LSTM, Transformers are not inherently capable of learning sequence information[18]. We first introduce vanilla positional encoding [17] to capture this information. If real-world timestamps are available, we use a learnable embedding layer to encode these timestamps [19].

https://doi.org/10.52202/079017-2982

Combining positional encoding with timestamp encoding, for any node $v \in \mathcal{V}$, the input to the temporal sparse attention module can be formulated as follows.

$$\begin{aligned}
&\overline{\boldsymbol{X}}_v = \boldsymbol{X}_v + PE(\boldsymbol{X}_v) + MLP(\boldsymbol{U}),\\
&PE_{pos,2i} = sin\left(pos/10000^{2i/d_{model}}\right)\\
&PE_{pos,2i+1} = cos\left(pos/10000^{2i/d_{model}}\right)
\end{aligned}$$

where $\boldsymbol{X}_v$ is associated time-series of node v; $\boldsymbol{U}$ is the available real-world time information, such as the hour of the day.

For any node v, $\overline{\boldsymbol{X}}_v$ serves as the key, query, and value in the temporal self-attention mechanism. For node v, similar to the spatial module, we denote the available time steps by $\tilde{\mathcal{T}}_v$. Then the temporal attention score and the temporal aggregation are formulated as:

$$\begin{aligned}
&\boldsymbol{Q}_v^t, \boldsymbol{K}_v^t, \boldsymbol{V}_v^t = \overline{\boldsymbol{X}}_v \boldsymbol{W}_Q^t, \overline{\boldsymbol{X}}_v \boldsymbol{W}_K^t, \overline{\boldsymbol{X}}_v \boldsymbol{W}_V^t \\
&(\alpha_{t_1,t_2}^t)_v = \frac{\exp(\langle \boldsymbol{q}_{v,t_1}^t, \boldsymbol{k}_{v,t_2}^t \rangle)}{\sum_{t_2' \in \tilde{\mathcal{T}}_v} \exp(\langle \boldsymbol{q}_{v,t_1}^t, \boldsymbol{k}_{v,t_2'}^t \rangle)},\\
&\boldsymbol{h}_{v,t}^t = \text{MLP}\left(\sum\nolimits_{t' \in \tilde{\mathcal{T}}_v} (\alpha_{t,t'}^t)_v \boldsymbol{v}_{v,t'}^t\right),
\end{aligned} \tag{6}$$

where $\boldsymbol{Q}_v^t, \boldsymbol{K}_v^t, \boldsymbol{V}_v^t \in \mathbb{R}^{T \times F}$ represent the query, key, value of spatial self-attention; $\boldsymbol{q}_{v,i}^t$, $\boldsymbol{k}_{v,i}^t$ and $\boldsymbol{v}_{v,i}^t$ stand for their i-the row; $\boldsymbol{W}_Q^t, \boldsymbol{W}_K^t, \boldsymbol{W}_V^t \in \mathbb{R}^{F \times F}$ are parameter matrices.

Based on these two sparse attention modules, we propagate all available information to the global missing data in one step. We also give a PAC-Learnability analysis for the proposed method.

Proposition 5.1. *Let $\mathcal{F}$ be a sparse attention-based imputation model class. If we draw a sample $\tilde{S}$ of size m, for any $f \in \mathcal{F}$, the following inequality holds.*

$$\mathbb{P}_{\tilde{S} \sim \tilde{\mathcal{D}}^m}\left[R_{\tilde{\mathcal{D}}}(h_{\tilde{S}}) \leq \epsilon\right] \geq 1 - \delta$$

with $m \geq \frac{\log 1/\delta}{2[\epsilon - 4\rho\mathcal{C} - 2C_l \Re_m(\mathcal{F})]}$, where $\mathcal{C} = C_l \cdot C_\phi \cdot B_d \cdot B_v$.

Remark 5.2. It can be seen that the number of required samples is positively correlated with the sparsity ρ. Unlike iteration-based models, sparsity is the sole factor that constrains the performance of the sparse attention-based model when dealing with incomplete data. Consequently, the proposed one-step strategy is scalable to various spatio-temporal data of different sizes and structures.

5.2 Confidence-based Iterative Refinement

In the first stage, we reconstructed the missing representations based on inherent spatial and temporal information. However, these two independent modules decompose the global context. Integrating these results and refining them through spatio-temporal structure is a straightforward idea. To further propagate reliable information, we propose a confidence-based message-passing mechanism. Intuitively, there are correlations between any pairwise spatio-temporal points. Thus, the missing data will remove some dependencies. The larger the correlations between an ST point and other observations, the more plausible its recovery. Therefore, for any ST point (u, t), we can define the confidences of its spatial and temporal imputation results as:

$$\beta_{u,t} = \frac{\sum_{v \in \tilde{\mathcal{V}}} \exp(\langle \boldsymbol{q}_u^s, \boldsymbol{k}_v^s \rangle)}{\sum_{v \in \mathcal{V}} \exp(\langle \boldsymbol{q}_u^s, \boldsymbol{k}_v^s \rangle)} + \frac{\sum_{t' \in \tilde{\mathcal{T}}_v} \exp(\langle \boldsymbol{q}_t^t, \boldsymbol{k}_{t'}^t \rangle)}{\sum_{t' \in \mathcal{T}} \exp(\langle \boldsymbol{q}_t^t, \boldsymbol{k}_{t'}^t \rangle)} \tag{7}$$

We extract highly confident parts from separate imputation results in temporal and spatial dimensions. However, simple weighting ignores the global context in spatio-temporal data. Therefore, we propose to employ the learned confidences as propagation weights through the spatio-temporal dimension. For any ST point (u, t), we first define the set of its neighbors as

$$\mathcal{N}_{u,t} = \{(v, t) | v \in \mathcal{N}_u\} \cup \{(v, t') | t' \in \mathcal{T} \backslash t\},$$

where $\mathcal{N}_u$ denotes the set of neighboring nodes of node u in the fixed graph G. Then the layer-wise message-passing for ST point (u, t) can be formulated as

$$\boldsymbol{O}_{u,t}^{l+1} = \text{MLP}\left(\boldsymbol{O}_{u,t}^{l} \| \sum\nolimits_{(v,t')\in\mathcal{N}_{u,t}} \beta_{v,t'} \cdot \boldsymbol{O}_{v,t'}^{l}\right),$$

where $\boldsymbol{O}_{u,t}^{0} = \boldsymbol{h}_{u,t}^{s} + \boldsymbol{h}_{u,t}^{t}$. During L layers, we finally recover representations for all ST points. Then the decoder is designed depend on different downstream tasks.

6 Experiments

We evaluate our approach on three sets of real-world datasets for different downstream tasks, including imputation and prediction. More experimental details and time complexity analysis are provided in the appendix. The source code and datasets are available at `https://github.com/dleyan/OPCR`.

6.1 Datasets

We consider three sets of spatio-temporal datasets and summarize their statistics in Table 1. It is evident that these datasets vary in terms of topology size. The T4C22 dataset primarily comprises spatially missing data, exhibiting an exceptionally high sparsity of up to 90%.

Table 1: Statistics of the datasets.

	TRAFFIC		LARGE-SCALE		T4C22		
	PEMS-BAY	METR-LA	PV-US	CER-E	LONDON	MADRID	MELBOURNE
# NODES	325	207	5166	6435	59110	63397	49510
# EDGES	2369	1515	71446	51428	132414	121902	94871
# STEPS	52128	34272	8688	8688	10560	10464	10176

To simulate incomplete data in realistic scenario, we design two policies to inject missing data into original datasets: 1) `Point Missing`, in which we follow the same setup of [12, 11], randomly dropping ρ of the available data. 2) `Spatial missing`, in which we are inspired by the T4C22 dataset, randomly dropping 25% of the available data, then mask out ρ of the devices.

6.2 Baselines

We compare the proposed OPCR with various baselines designed for different data types.

- **Tabular Imputation.** We consider three statistical methods for matrix imputation: 1) `Mean`, imputing missing data using the mean value at each time step. 2) `Matrix Factorization (MF)` [20] with rank = 10. 3) `GRAPE` [21], a graph-based feature imputation method, which regards observations and features as two types of nodes in a bipartite graph, and the observed feature values as edges.
- **Graph Imputation.** For spatial missing, we consider some node attributes imputation methods: 1) `Feature Propagation (FP)` [3], iteratively propagating observed messages through graph structure. 2) `PCFI` [4], further measuring the propagation confidence.
- **Time-series Imputation.** To impute multivariate time-series, we consider: 1) `BRITS` [7], a bidirectional RNN-based model. 2) `SAITS` [8], a transformer-based model. 2) `CSDI` [9], a diffusion-based model.
- **Spatio-temporal Data Imputation**. We also consider some state-of-the-art methods for spatio-temporal data imputation. 1) `IGNNK` [10], an inductive GNN-based model. 3) `SPIN-H` [12], an efficient version of spatio-temporal attention based method. 3) `PriSTI` [13], a conditional diffusion model. 4) `PoGeVon` [14] , solving a specific issue in which spatio-temporal data contains missing values in both node time series features and graph structures.

6.3 Spatio-temporal Data Imputation Task

For the imputation task, we consider mean absolute error (MAE) as evaluation metrics. All the experiments are run with 5 different random seeds. In each round, we inject ρ of missing data into

https://doi.org/10.52202/079017-2982

Table 2: Imputation performances (in terms of MAE) on Traffic dataset and Large-scale dataset .

Model	Point Missing				Spatial Missing			
	PEMS-BAY	METR-LA	PV-US	CER-E	PEMS-BAY	METR-LA	PV-US	CER-E
Mean	5.00 ± 0.00	10.28 ± 0.02	8.34 ± 0.00	0.56 ± 0.00	5.08 ± 0.13	10.37 ± 0.21	8.45 ± 0.28	0.56 ± 0.01
MF	5.25 ± 0.00	7.55 ± 0.04	4.53 ± 0.01	0.50 ± 0.00	5.71 ± 0.07	7.61 ± 0.11	12.47 ± 0.24	0.57 ± 0.01
GRAPE	4.12 ± 0.02	6.75 ± 0.03	10.69 ± 0.22	0.34 ± 0.00	4.72 ± 0.02	6.78 ± 0.03	11.67 ± 0.03	0.36 ± 0.00
FP	4.92 ± 0.00	9.22 ± 0.02	7.60 ± 0.01	0.52 ± 0.00	5.05 ± 0.05	9.25 ± 0.16	7.87 ± 0.13	0.52 ± 0.02
PCFI	5.14 ± 0.01	12.94 ± 0.03	8.79 ± 0.02	0.52 ± 0.01	6.75 ± 0.59	14.95 ± 1.86	12.31 ± 0.30	0.50 ± 0.04
BRITS	2.58 ± 0.00	5.25 ± 0.10	30.48 ± 4.75	1.38 ± 0.99	5.64 ± 0.16	8.88 ± 0.52	25.77 ± 9.00	1.15 ± 0.28
SAITS	2.44 ± 0.01	5.21 ± 0.03	6.40 ± 4.56	0.67 ± 0.57	6.14 ± 0.24	9.51 ± 0.63	2.73 ± 0.08	0.37 ± 0.11
CSDI	2.16 ± 0.04	3.48 ± 0.09	7.51 ± 2.30	0.49 ± 0.02	3.76 ± 0.75	4.51 ± 0.45	8.97 ± 1.93	0.47 ± 0.02
IGNNK	2.61 ± 0.01	4.37 ± 0.02	7.86 ± 0.02	0.38 ± 0.00	4.70 ± 0.13	6.85 ± 0.27	11.43 ± 0.06	0.46 ± 0.00
SPIN-H	1.84 ± 0.02	2.99 ± 0.03	1.94 ± 0.06	0.33 ± 0.09	4.93 ± 0.04	7.62 ± 0.06	2.63 ± 0.06	0.29 ± 0.00
PoGeVon	5.68 ± 0.01	8.86 ± 0.01	/	/	5.73 ± 0.14	8.82 ± 0.56	/	/
PriSTI	2.05 ± 0.02	3.85 ± 0.18	11.93 ± 2.84	0.63 ± 0.12	3.05 ± 0.20	5.04 ± 0.75	11.75 ± 7.54	0.59 ± 0.14
OPCR	$\mathbf{1.79 \pm 0.01}$	$\mathbf{2.80 \pm 0.02}$	$\mathbf{1.73 \pm 0.06}$	$\mathbf{0.28 \pm 0.00}$	$\mathbf{2.27 \pm 0.07}$	$\mathbf{3.20 \pm 0.10}$	$\mathbf{2.01 \pm 0.11}$	$\mathbf{0.28 \pm 0.00}$

the entire dataset using different masking policies. All models are trained on the sparse training set and evaluated on the corresponding sparse testing set.

6.3.1 Highly Sparse Data Imputation

The experimental results of all models are summarized in Table 2, where the missing rate ρ is set as 95%. There is a clear gap in the imputation performance on the traffic dataset between spatial-missing data and point-missing data, indicating that recovering spatial-missing data is a challenging issue. For the tabular and graph imputation baselines, their performance remains relatively consistent across datasets with spatial and point missing. This can be easily understood, as these baselines only utilize or fail to utilize spatial structure. The time-series imputation baselines are not suitable for spatial missing, as they solely rely on temporal information. In addition, it is difficult for them to model such a high-dimensional time series on large-scale datasets.

For state-of-the-art (SOTA) spatio-temporal series imputation baseline SPIN-H, they are most suitable for small-scale data with point missing. Since they only propagate partial observation iteratively, larger topology and higher spatial missing rate imply a need for more iterations. Therefore, the error accumulation and information loss during iterative process lead to their poor performance. For the inductive baseline IGNNK, it converts the original task to smaller-scale spatio-temporal series via sampling sub-graphs. The effectiveness of IGNNK proves that there are correlations between all nodes, which is consistent with our motivation. However, this sampling strategy still ignores some of the valid information. Therefore, considering the problem of error accumulation in autoregressive methods, some diffusion-based models have been proposed, such as CSDI [9] and PriSTI [13]. It is still challenging for CSDI to model a high-dimensional time series in the large-scale dataset. However, it show performance improvement on the traffic dataset compared to other temporal baselines (i.e., BRITS and SAITS). For the spatio-temporal method PriSTI, it achieves better performances than SPIN-H on the traffic dataset with spatial missing. However, its performances are unstable on the large-scale datasets, especially the spatially missing data. PoGeVon [14] is proposed to deal with specific data where attribute-missing and structure-missing co-exist. Therefore, it show worse imputation performances than other spatio-temporal baselines. In addition, PoGeVon does not apply to the large-scale dataset because of its high computational complexity.

As a result, our proposed OPCR outperforms all the baselines in most cases. Especially, on traffic dataset with spatial missing, OPCR exceeds the best baseline by 26% and 37%. on large-scale PVUS dataset, OPCR exceeds the best baseline by 11% and 24% under two settings. These experiments demonstrate the effectiveness and scalability of our proposed OPCR.

6.3.2 Imputation with Increasing Missing Rate

To evaluate the effect of data sparsity on imputation performance, we simulate sparse data at various rate, ranging from 25% to 95%. Figures 2 and 3 report imputation performances on point missing data and spatially missing data, respectively. On the traffic dataset, it is clear that the imputation errors increase with the missing rate. Similar to the results in Table 2, spatial missing poses a greater

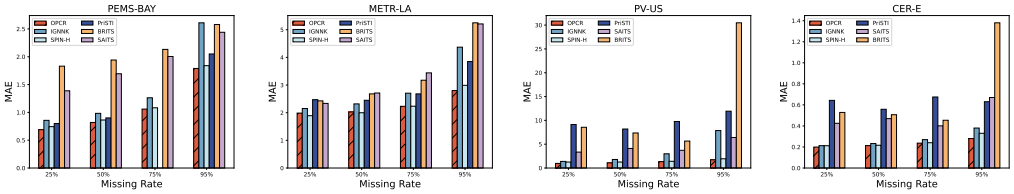

Figure 2: Imputation performance (in terms of MAE) with increasing point missing rate.

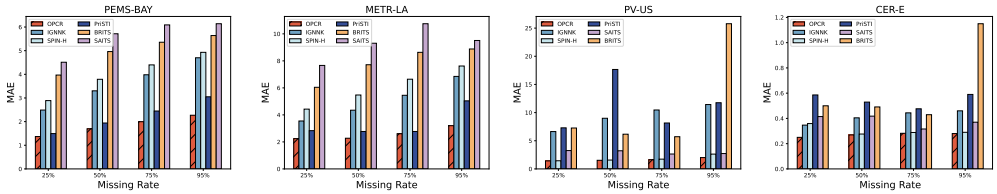

Figure 3: Imputation performance (in terms of MAE) with increasing saptial missing rate.

challenge for all models. On large-scale datasets, temporal baselines struggle to model extremely high-dimensional data and do not exhibit the same trends as observed in the traffic dataset. In addition, these models usually terminates training process early then result extremely poor imputation performances. For the inductive spatio-temporal baseline IGNNK, its overall performance show a significant improvement when the missing rate is reduced from 95% to 75%. This empirical results are due to IGNNK can not sample enough sub-graph data for training in highly-sparse data. For SPIN-H, its imputation error on the CER-E dataset is not exactly positively correlated with the missing rate, which is consistent with our theoretical findings. The diffusion-based method PriSTI achieves significant improvements over other baselines on the traffic dataset with spatial missing. However, PriSTI struggle to deal with large-scale dataset. For the proposed OPCR, we can observe that it outperforms baselines in most cases. Furthermore, the imputation performances of our method steadily improves with decreasing missing rates. This experimental result is consistent with the theoretical findings, confirming that the PAC learnability of the proposed OPCR is only constrained by data sparsity.

6.4 Traffic Prediction Task

For the T4C22 dataset, we follow the competition rules. We set the evaluation metric to a weighted cross-entropy loss for the congestion classification task, and MAE for the prediction task. All the models are equipped with the same downstream network and use the same training strategy. Table 3 presents the traffic performance comparison for two downstream tasks. Since we designed strong downstream models for both prediction tasks, the performance gap is insignificant. It can be seen that our proposed OPCR outperforms baselines across all experiments. There are also some findings in the traffic prediction experiments. SPIN-H performs poorly compared to graph imputation baselines (i.e., FP and PCFI). This is because the T4C22 dataset only considers short-term (i.e., 4 time steps) historical series, and its highly spatial sparsity results in slow propagation of valid information.

Table 3: Traffic Prediction performances.

Model	Congestion Classification			Travel Time Prediction		
	LONDON	MADRID	MELBOURNE	LONDON	MADRID	MELBOURNE
FP	0.8126	0.8624	0.8635	84.56	68.17	31.48
PCFI	0.8159	0.8514	0.8552	105.61	61.06	31.43
SPIN-H	0.8335	0.8768	0.8795	77.21	59.24	**31.36**
OPCR	**0.8114**	**0.8444**	**0.8519**	**74.44**	**55.39**	31.54

6.5 Ablation Study

To evaluate the effectiveness of different components of our proposed OPCR, we make comparisons between some model variants. To demonstrate the robustness of our model, we also conduct sensitivity

https://doi.org/10.52202/079017-2982

analysis with respect to model's hyper-parameters. All models here are trained on the traffic datasets and large-scale datasets with 95% missing rate.

Effectiveness of Each Component. We consider the following variants. 1) `Spatial module (S)`: Only use spatial module as encoder. 2) `Temporal module (T)`: Only use temporal module as encoder. 3) `W/O refinement (W/O R)`: Ignore the confidence-based refinement.

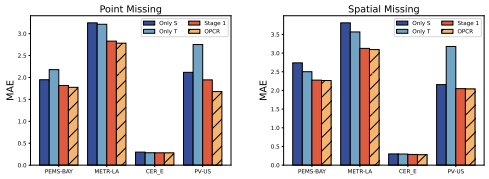

Figure 4: The effect of each component.

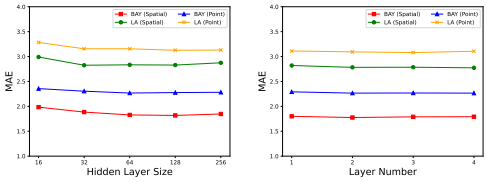

Figure 5: The effect of hyper-parameters.

The results of the ablation study are shown in Fig. 4. First, we compare the spatial module and temporal module. Obviously, it is challenging for both variants to handle spatially missing data. On the large-scale dataset, the temporal module presents poorer performances compared to the spatial module. This result is easy to understand because the temporal module utilizes a narrower range of valid information. Second, combining temporal-based and spatial-based results yields superior performances compared to a single module. Finally, when comparing the complete model with all variants, it can be seen that confidence-based propagation also plays a vital role.

Sensitivity w.r.t. Hyper-parameters. First, we consider the effect of hidden states, as shown in Fig. 5. As the hidden states go from 16 to 256, the overall performance slowly increases and then fluctuates slightly. Therefore, we set the number of hidden states to 128. This result reflects the robustness of the proposed model. Then, we consider the sensitivity w.r.t. the number of layers in stage 2, as shown in Fig. 5. The approximate best level can be reached when the number of layers is only 2. These experiments prove that our design motivation is that a shallow iterative propagation is enough to refine the complete recovery by the first stage.

7 Conclusion

For the highly sparse spatio-temporal data learning problem prevalent in the real-world, the existing common scheme is iterative propagation-base imputation. However, such methods empirically suffer from error accumulation and high computational costs in large-scale data. Therefore, we first provide a PAC-learnability analysis for iterative imputation methods from the perspective of PAC-learnability, which theoretically demonstrates their limitations. Motivated by theoretical findings, we propose one-step propagation and confidence-based refinement (OPCR). In the first stage, we rapidly propagate useful information to all missing data through a sparse attention mechanism that makes full use of limited observations. To eliminate the bias caused by independent temporal and spatial modeling, we propose assigning confidence to imputation results and achieving more accurate spatio-temporal message-passing. We evaluate the imputation and prediction performances of the proposed OPCR on a dataset of different scales. The experimental results illustrate our approach outperforms the state-of-the-art imputation methods. There are several interesting future directions. First, it would be interesting to explore some downstream tasks without supervision. The second is to extend our theoretical analysis to non-random missing scenarios.

8 Acknowledgment

The work was supported by grants from the National Natural Science Foundation of China (No. 92367110).

References

[1] Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys (CSUR)*, 51(4):1–41, 2018.

[2] Moritz Neun, Christian Eichenberger, Henry Martin, Markus Spanring, Rahul Siripurapu, Daniel Springer, Leyan Deng, Chenwang Wu, Defu Lian, Min Zhou, et al. Traffic4cast at neurips 2022–predict dynamics along graph edges from sparse node data: Whole city traffic and eta from stationary vehicle detectors. In *NeurIPS 2022 Competition Track*, pages 251–278. PMLR, 2022.

[3] Emanuele Rossi, Henry Kenlay, Maria I Gorinova, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M Bronstein. On the unreasonable effectiveness of feature propagation in learning on graphs with missing node features. In *Learning on Graphs Conference*, pages 11–1. PMLR, 2022.

[4] Daeho Um, Jiwoong Park, Seulki Park, and Jin Young Choi. Confidence-based feature imputation for graphs with partially known features. *arXiv preprint arXiv:2305.16618*, 2023.

[5] Xu Chen, Siheng Chen, Jiangchao Yao, Huangjie Zheng, Ya Zhang, and Ivor W Tsang. Learning on attribute-missing graphs. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):740–757, 2020.

[6] Wenxuan Tu, Sihang Zhou, Xinwang Liu, Yue Liu, Zhiping Cai, En Zhu, Zhang Changwang, and Jieren Cheng. Initializing then refining: A simple graph attribute imputation network. In *Proc. IJCAI*, pages 3494–3500, 2022.

[7] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018.

[8] Wenjie Du, David Côté, and Yan Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023.

[9] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, 34:24804–24816, 2021.

[10] Yuankai Wu, Dingyi Zhuang, Aurelie Labbe, and Lijun Sun. Inductive graph neural networks for spatiotemporal kriging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4478–4485, 2021.

[11] Cini Andrea, Marisca Ivan, Cesare Alippi, et al. Filling the g_ap_s: Multivariate time series imputation by graph neural networks. In *ICLR 2022*, pages 1–20. 2021.

[12] Ivan Marisca, Andrea Cini, and Cesare Alippi. Learning to reconstruct missing data from spatiotemporal graphs with sparse observations. *Advances in Neural Information Processing Systems*, 35:32069–32082, 2022.

[13] Mingzhe Liu, Han Huang, Hao Feng, Leilei Sun, Bowen Du, and Yanjie Fu. Pristi: A conditional diffusion framework for spatiotemporal imputation. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 1927–1939. IEEE, 2023.

[14] Dingsu Wang, Yuchen Yan, Ruizhong Qiu, Yada Zhu, Kaiyu Guan, Andrew J Margenot, and Hanghang Tong. Networked time series imputation via position-aware graph enhanced variational autoencoders. *arXiv preprint arXiv:2305.18612*, 2023.

[15] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[16] Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4602–4609, 2019.

 https://doi.org/10.52202/079017-2982

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[18] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*, 2022.

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

[20] Andrzej Cichocki and Anh-Huy Phan. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 92(3):708–721, 2009.

[21] Jiaxuan You, Xiaobai Ma, Yi Ding, Mykel J Kochenderfer, and Jure Leskovec. Handling missing data with graph representation learning. *Advances in Neural Information Processing Systems*, 33:19075–19087, 2020.

[22] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

[23] Andrea Cini, Ivan Marisca, Filippo Maria Bianchi, and Cesare Alippi. Scalable spatiotemporal graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 7218–7226, 2023.

[24] Moritz Neun, Christian Eichenberger, Henry Martin, Markus Spanring, Rahul Siripurapu, Daniel Springer, Leyan Deng, Chenwang Wu, Defu Lian, Min Zhou, et al. Traffic4cast at neurips 2022–predict dynamics along graph edges from sparse node data: Whole city traffic and eta from stationary vehicle detectors. *arXiv preprint arXiv:2303.07758*, 2023.

[25] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*, 2019.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[27] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[28] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[29] Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

A Experiment

A.1 Datasets

We consider the following three sets of spatio-temporal datasets.

- **Traffic dataset**[22] record traffic dynamics every 5 minutes in San Francisco Bay Area and LA County, respectively [11].
- **Large-scale dataset**[23] includes two larger-scale datasets. The sources of the PV-US dataset and the CER-E dataset are the simulated energy production by PV farms [1] and the CER Smart Metering Project [2], respectively. Both datasets are aggregated in 30 minutes.
- **T4C22 dataset** [24] record traffic dynamics every 15 minutes from vehicle detectors in three cities, which is provided by Traffic4cast 2022 competition [3].

For the traffic dataset and PV-US dataset, we derive the adjacency from pairwise geographic distances [22, 25]. Without location information, the adjacency of the CER-E dataset is derived from similarity between historical series. For the T4C22 dataset, the available urban road networks can serve as adjacency.

A.2 Experimental settings

Considering memory consumption, for SPIN, we choose the efficient version SPIN-H as baseline. On the traffic dataset, we train all models with an RTX 3090 GPU (24GB RAM). On the large-scale dataset and the T4C22 dataset, we train them with a V100 GPU (16GB RAM). All the baselines have been implemented in PyTorch [26]. We use, whenever possible, the open-source code and configuration provided by the authors.

For spatio-temporal data imputation task, we select windows of length 24. On the traffic datasets, we strictly follow the settings of SPIN[4]. We fix the maximum number of epochs to 300 and we use early stopping on the validation set with patience of 40 epochs. On the large-scale dataset, considering memory capacity and computational efficiency, we reduced the number of hidden states to 16 for some baselines (i.e., SPIN-H, CSDI, and PriSTI). We fix the maximum number of epochs to 200 and we use early stopping on the validation set with patience of 10 epochs.

For traffic prediction task, we follow the competition settings. Since the original test dataset is unlabeled, we randomly select data from the training set for different days. Note that this partitioning strategy is aligns with the competition rules. We provided each model with the same downstream network and trained the models end-to-end. For congestion classification task, we train all the models for 20 epochs. For travel time prediction task, we train all the models for 50 epochs. We both select the best model with the minimum loss on the validation set for evaluation.

A.3 Time Complexity and Run-time Analysis

We compare our proposed OPCR with SOTA iterative imputation baselines SPIN. We let the number of edges in G be E and the number of iterations for all models be L. According to [12], the time complexity of SPIN scales with $\mathcal{O}\left(L(N+E)T^2\right)$ and its efficient version SPIN-H scales with $\mathcal{O}\left(L(N+E)KT\right)$, where $K \ll T$.

For our proposed OPCR, in the first stage, we propagate observations to missing data, thus we can reduce the time complexity from $\mathcal{O}\left(N^2+NT^2\right)$ to $\mathcal{O}\left((N^2+NT^2)\times\rho(1-\rho)\right)$ via sparse attention matrices. Considering the sparsity of the graph structure, it is reasonable to assume that E and $N^2\times\rho(1-\rho)$ are similar in magnitude. Moreover, since we learn spatial attention matrix from static node features and share it across all spatio-temporal series within a batch, the the computation remains efficient despite the introduction of attention-based confidence assignment. In the second stage, the time complexity of layer-wise updation is $\mathcal{O}\left(ET+NT^2\right)$. As we mentioned

[1] https://www.nrel.gov/grid/solar-power-data.html

[2] https://www.ucd.ie/issda/data/commissionforenergyregulationcer/

[3] https://www.iarai.ac.at/traffic4cast/challenge/

[4] `https://github.com/Graph-Machine-Learning-Group/spin`

https://doi.org/10.52202/079017-2982

in Section 4, after one-step propagation, OPCR requires fewer iterations for refinement compared to SPIN. Therefore, our proposed method actually has greater computational efficiency and scalability.

In addition, following [23], we present the timings and memory consumption in the Table 4. Compared to the spatio-temporal imputation baselines SPIN-H, our proposed model shows significant superiority in computational efficiency and memory usage.

Table 4: Timings and memory consumption.

MODEL	PV-US			CER-E		
	MEMORY	BATCH/S	BATCH SIZE	MEMORY	BATCH/S	BATCH SIZE
BRITS	11.20GB	2.84	32	16.53GB	1.97	32
SAITS	4.52GB	8.95	128	5.28GB	1.76	128
SPIN-H	14.22GB	2.88	2	18.00GB	3.46	2
PRISTI	14.22GB	1.40	4	17.52GB	1.01	4
OPCR	16.98GB	3.09	4	18.22GB	2.87	4

B Proof of Proposition4.3

Proposition B.1. *Assume that all ST points are randomly masked with a probability of ρ. Let $\mathcal{F}$ be a K-iterations imputation model class. If we draw a sample $\tilde{S}$ of size m, for any $f \in \mathcal{F}$, the following inequality holds.*

$$\mathbb{P}_{\tilde{S}\sim\tilde{\mathcal{D}}^m}\left[R_{\tilde{\mathcal{D}}}(h_{\tilde{S}}) \leq \epsilon\right] \geq 1-\delta$$

with $m \geq \frac{\log 1/\delta}{2\epsilon - 4\eta\mathcal{C}K - 4(1-\eta)\mathcal{C}K\cdot\rho^{\tau} - 4C_l\Re_m(\mathcal{F})}$.

Given an sparse training set $\tilde{S} \sim \tilde{\mathcal{D}}^m$, for any $f \in \mathcal{F}$, we have

$$R_{\tilde{\mathcal{D}}}(f) - \hat{R}_{\tilde{S}}(f) = \left[R_{\tilde{\mathcal{D}}}(f) - R_{\mathcal{D}}(f)\right] + \left[R_{\mathcal{D}}(f) - \hat{R}_{S}(f)\right] + \left[\hat{R}_{S}(f) - \hat{R}_{\tilde{S}}(f)\right] \tag{8}$$

Corollary B.2. *For any $f \in \mathcal{F}$ and any $(\boldsymbol{X}, \boldsymbol{Y}) \in \mathcal{X}\times\mathcal{Y}$, we mask all ST points at random with probability of ρ. We denote this mask distribution as $\mathcal{M}$. Given a complete spatio-temporal series $\boldsymbol{X}$ and its sparse version $\tilde{\boldsymbol{X}}$, we denote the node-level output of the model as $\hat{\boldsymbol{Y}} = f(\boldsymbol{X})$ and $\tilde{\boldsymbol{Y}} = f(\tilde{\boldsymbol{X}})$. The difference between the model's prediction error on complete data and sparse data have the following upper bound.*

$$\mathbb{E}_{\boldsymbol{M}\sim\mathcal{M}}\left|L(\tilde{\boldsymbol{Y}}, \boldsymbol{Y}) - L(\hat{\boldsymbol{Y}}, \boldsymbol{Y})\right| \leq \mathcal{C}\cdot\sum_{k=1}^{K}\left[\rho^{\tau_{\min}^{k-1}} + \eta\cdot(1-\rho^{\tau_{\max}^{k-1}})\right],$$

where $\mathcal{C} = C_l\cdot\gamma^K\cdot C_{\phi}^{K+1}\cdot(d_{\max})^K\cdot B_X B_d$.

Proof. Under assumptions, f is an encoder-decoder architecture, for any complete samples $(\boldsymbol{X}, \boldsymbol{Y})$ and its sparse version $(\tilde{\boldsymbol{X}}, \boldsymbol{Y})$, we have

$$\begin{aligned}
\left|L(\tilde{\boldsymbol{Y}}, \boldsymbol{Y}) - L(\hat{\boldsymbol{Y}}, \boldsymbol{Y})\right| &= \left|\frac{1}{N}\sum_{v\in\mathcal{V}}\left(l\left(\tilde{\boldsymbol{y}}_v, \boldsymbol{y}_v\right) - l\left(\hat{\boldsymbol{y}}_v, \boldsymbol{y}_v\right)\right)\right| \\
&\leq C_l\cdot\frac{1}{N}\sum_{v\in\mathcal{V}}\left\|\tilde{\boldsymbol{y}}_v - \hat{\boldsymbol{y}}_v\right\|_2 \\
&= C_l\cdot\frac{1}{N}\sum_{v\in\mathcal{V}}\left\|f_d(f_e(\tilde{\boldsymbol{X}}))_v - f_d(f_e(\boldsymbol{X}))_v\right\|_2.
\end{aligned} \tag{9}$$

Here Eq.9 is due to the loss function l is C_l-lipschitz continuous. As the assumption about decoder f_d, we have

$$\begin{aligned}
\left|L(\tilde{\boldsymbol{Y}}, \boldsymbol{Y}) - L(\hat{\boldsymbol{Y}}, \boldsymbol{Y})\right| &\leq C_l \cdot \frac{1}{N} \sum_{v \in \mathcal{V}} \left\| f_d(f_e(\tilde{\boldsymbol{X}}))_v - f_d(f_e(\boldsymbol{X}))_v \right\|_2 \\
&= C_l \cdot \frac{1}{N} \sum_{v \in \mathcal{V}} \left\| \phi\left(\frac{1}{T} \sum_{t \in \mathcal{T}} f_e(\tilde{\boldsymbol{X}})_{v,t} \boldsymbol{W}_d\right) - \phi\left(\frac{1}{T} \sum_{t \in \mathcal{T}} f_e(\boldsymbol{X})_{v,t} \boldsymbol{W}_d\right) \right\|_2 \\
&\leq C_l \cdot C_\phi \cdot \frac{1}{NT} \sum_{v \in \mathcal{V}} \left\| \sum_{t \in \mathcal{T}} \left(f_e(\tilde{\boldsymbol{X}})_{v,t} - f_e(\boldsymbol{X})_{v,t}\right) \boldsymbol{W}_d \right\|_2 \qquad (10) \\
&\leq C_l \cdot C_\phi \cdot \frac{1}{NT} \sum_{v \in \mathcal{V}} \sum_{t \in \mathcal{T}} \left\| f_e(\tilde{\boldsymbol{X}})_{v,t} - f_e(\boldsymbol{X})_{v,t} \right\|_2 \cdot \|\boldsymbol{W}_d\|_2 \\
&\leq C_l \cdot C_\phi \cdot B_d \cdot \frac{1}{NT} \sum_{v \in \mathcal{V}} \sum_{t \in \mathcal{T}} \left\| f_e(\tilde{\boldsymbol{X}})_{v,t} - f_e(\boldsymbol{X})_{v,t} \right\|_2 . \qquad (11)
\end{aligned}$$

Here Eq.10 is due to the ϕ is C_ϕ-lipschitz continous and Eq.11 is due to $\|\boldsymbol{W}_d\|_2 \leq B_d$. Since the encoder f_e is a K-iteration imputation model, we denote the ST point representations after k iterations by $\boldsymbol{H}^k$ and $\tilde{\boldsymbol{H}}^k$. We denote the set of recovered ST points after k iterations by $\mathcal{N}_o^k$. For the number of iterations K, we assume that K exactly satisfies $\mathcal{Z}_o^K = \mathcal{Z}$ and for any $\mathcal{Z}_o^{K-1} \subset \mathcal{Z}$. Thus, we have

$$\begin{aligned}
\frac{1}{NT} \sum_{v \in \mathcal{V}} \sum_{t \in \mathcal{T}} \left\| f_e(\tilde{\boldsymbol{X}})_{v,t} - f_e(\boldsymbol{X})_{v,t} \right\|_2 &= \frac{1}{NT} \cdot \sum_{v \in \mathcal{V}} \sum_{t \in \mathcal{T}} \left\| \tilde{\boldsymbol{h}}_{v,t}^K - \boldsymbol{h}_{v,t}^K \right\|_2 \\
&= \frac{1}{NT} \cdot \sum_{z \in \mathcal{Z}} \left\| \tilde{\boldsymbol{h}}_z^K - \boldsymbol{h}_z^K \right\|_2 , \qquad (12)
\end{aligned}$$

where $\mathcal{Z} = \{(v,t) | v \in \mathcal{V}, t \in \mathcal{T}\}$. Then we have the following recursion:

$$\begin{aligned}
&\sum_{z \in \mathcal{Z}_o^k} \left\| \tilde{\boldsymbol{h}}_z^k - \boldsymbol{h}_z^k \right\|_2 \\
&= \sum_{z \in \mathcal{Z}_o^k} \left\| \phi\left(\sum_{z' \in \tilde{\mathcal{N}}_z^k} \tilde{a}_{z' \to z}^k \tilde{\boldsymbol{h}}_{z'}^{k-1}\right) - \phi\left(\sum_{z' \in \mathcal{N}_z} a_{z' \to z} \boldsymbol{h}_{z'}^{k-1}\right) \right\|_2 \\
&\leq C_\phi \cdot \sum_{z \in \mathcal{Z}_o^k} \left\| \sum_{z' \in \tilde{\mathcal{N}}_z^k} \tilde{a}_{z' \to z} \tilde{\boldsymbol{h}}_{z'}^{k-1} - \sum_{z' \in \mathcal{N}_z} a_{z' \to z} \boldsymbol{h}_{z'}^{k-1} \right\|_2 \qquad (13) \\
&\leq C_\phi \cdot \sum_{z \in \mathcal{Z}_o^k} \left(\left\| \sum_{z' \in \tilde{\mathcal{N}}_z^k} \tilde{a}_{z' \to z}^k \left(\tilde{\boldsymbol{h}}_{z'}^{k-1} - \boldsymbol{h}_{z'}^{k-1}\right) \right\|_2 + \left\| \sum_{z' \in \mathcal{N}_z} \left(\tilde{a}_{z' \to z}^k - a_{z' \to z}\right) \boldsymbol{h}_{z'}^{k-1} \right\|_2 \right) \qquad (14) \\
&\leq C_\phi \cdot \sum_{z \in \mathcal{Z}_o^k} \left(\sum_{z' \in \tilde{\mathcal{N}}_z^k} |\tilde{a}_{z' \to z}^k| \cdot \left\| \tilde{\boldsymbol{h}}_{z'}^{k-1} - \boldsymbol{h}_{z'}^{k-1} \right\|_2 + \sum_{z' \in \mathcal{N} \backslash \tilde{\mathcal{N}}_z^k} |a_{z' \to z}| \cdot \left\| \boldsymbol{h}_{z'}^{k-1} \right\|_2 + \sum_{z' \in \tilde{\mathcal{N}}_z^k} |\tilde{a}_{z' \to z}^k - a_{z' \to z}| \cdot \left\| \boldsymbol{h}_{z'}^{k-1} \right\|_2 \right)
\end{aligned}$$

Here Eq.13 is due to that ϕ is C_ϕ-lipschitz continuous. Eq.14 is due to $\tilde{a}_{z' \to z}^k = 0$ if $z' \in \mathcal{N}_z \backslash \tilde{\mathcal{N}}_z^k$.

For $\forall z \in \mathcal{Z}$ and $\forall z' \in \mathcal{N}_z$, we assume that $0 < a_{z' \to z} \leq \gamma$. For $\forall z \in \mathcal{Z}$, $\forall z' \in \mathcal{N}_z^k$ and $\forall k \in [1, K]$, we assume that $0 < \tilde{a}_{z' \to z}^k \leq \gamma$ and $0 < |a_{z' \to z} - \tilde{a}_{z' \to z}^k| \leq \gamma' < \gamma$. These assumptions are easy to implement because the model tends to use positive weights when aggregating neighboring ST points.

https://doi.org/10.52202/079017-2982

For convenience, let $0 < \eta = \frac{\gamma'}{\gamma} < 1$, we have

$$\begin{aligned}
&\sum_{z\in\mathcal{Z}_o^k} \left\|\tilde{\boldsymbol{h}}_z^k - \boldsymbol{h}_z^k\right\|_2 \\
&\leq C_\phi \cdot \sum_{z\in\mathcal{Z}_o^k} \left(\sum_{z'\in\tilde{\mathcal{N}}_z^k} |\tilde{a}_{z'\to z}^k| \cdot \left\|\tilde{\boldsymbol{h}}_{z'}^{k-1} - \boldsymbol{h}_{z'}^{k-1}\right\|_2 + \sum_{z'\in\mathcal{N}\setminus\tilde{\mathcal{N}}_z^k} |a_{z'\to z}| \cdot \left\|\boldsymbol{h}_{z'}^{k-1}\right\|_2 + \sum_{z'\in\tilde{\mathcal{N}}_z^k} |\tilde{a}_{z'\to z}^k - a_{z'\to z}| \cdot \left\|\boldsymbol{h}_{z'}^{k-1}\right\|_2 \right) \\
&\leq C_\phi \cdot \sum_{z\in\mathcal{Z}_o^k} \left(\sum_{z'\in\tilde{\mathcal{N}}_z^k} \gamma \cdot \left\|\tilde{\boldsymbol{h}}_{z'}^{k-1} - \boldsymbol{h}_{z'}^{k-1}\right\|_2 + \sum_{z'\in\mathcal{N}_z\setminus\tilde{\mathcal{N}}_z^k} \gamma \cdot \left\|\boldsymbol{h}_{z'}^{k-1}\right\|_2 + \sum_{z'\in\tilde{\mathcal{N}}_z^k} \gamma' \cdot \left\|\boldsymbol{h}_{z'}^{k-1}\right\|_2 \right) \qquad (15) \\
&\leq \gamma C_\phi d_{\max} \cdot \sum_{z\in\mathcal{Z}_o^{k-1}} \left\|\tilde{\boldsymbol{h}}_{z'}^{k-1} - \boldsymbol{h}_{z'}^{k-1}\right\|_2 + \gamma C_\phi \cdot \sum_{z\in\mathcal{Z}_o^k} \left(\sum_{z'\in\mathcal{N}_z\setminus\tilde{\mathcal{N}}_z^k} \left\|\boldsymbol{h}_{z'}^{k-1}\right\|_2 + \sum_{z'\in\tilde{\mathcal{N}}_z} \eta \cdot \left\|\boldsymbol{h}_{z'}^{k-1}\right\|_2 \right) \qquad (16)
\end{aligned}$$

For the second term and the third term in Eq. 16, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{M}} \left[\sum_{z\in\mathcal{Z}_o^k} \left(\sum_{z'\in\mathcal{N}_z\setminus\tilde{\mathcal{N}}_z} \left\|\boldsymbol{h}_{z'}^{k-1}\right\|_2 + \sum_{z'\in\tilde{\mathcal{N}}_z} \eta \cdot \left\|\boldsymbol{h}_{z'}^{k-1}\right\|_2 \right) \right] \\
&\leq \mathbb{E}_{\boldsymbol{M}} \left[\sum_{z\in\mathcal{Z}} \left(\sum_{z'\in\mathcal{N}_z\setminus\tilde{\mathcal{N}}_z} \left\|\boldsymbol{h}_{z'}^{k-1}\right\|_2 + \sum_{z'\in\tilde{\mathcal{N}}_z} \eta \cdot \left\|\boldsymbol{h}_{z'}^{k-1}\right\|_2 \right) \right] \\
&\leq d_{\max} \cdot \left(\rho^{\tau_{\min}^{k-1}} \cdot \sum_{z\in\mathcal{Z}} \left\|\boldsymbol{h}_z^{k-1}\right\|_2 + \eta \cdot (1 - \rho^{\tau_{\max}^{k-1}}) \cdot \sum_{z\in\mathcal{Z}} \left\|\boldsymbol{h}_z^{k-1}\right\|_2 \right) \\
&= d_{\max} \cdot \left[\rho^{\tau_{\min}^{k-1}} + \eta \cdot (1 - \rho^{\tau_{\max}^{k-1}}) \right] \cdot \sum_{z\in\mathcal{Z}} \left\|\boldsymbol{h}_z^{k-1}\right\|_2 . \qquad (17)
\end{aligned}$$

Here $\tau_{\max}^k = \max_{z\in\mathcal{Z}} \tau_z^k$ and $\tau_{\min}^k = \min_{z\in\mathcal{Z}} \tau_z^k$, where $\tau_z^k = |\{z' | d_{z'\to z} \leq k, z' \in \mathcal{Z}_o^{k-1}\}|$. Actually, τ_z^k represents the number of k-hop neighbors of ST point z.

For $\sum_{z\in\mathcal{Z}} \left\|\boldsymbol{h}_z^k\right\|_2$, we have

$$\begin{aligned}
\sum_{z\in\mathcal{Z}} \left\|\boldsymbol{h}_z^k\right\|_2 &= \sum_{z\in\mathcal{Z}} \left\| \phi\left(\sum_{z'\in\mathcal{N}_z} a_{z'\to z} \boldsymbol{h}_{z'}^{k-1} \right) \right\|_2 \\
&= \sum_{z\in\mathcal{Z}} \left\| \phi\left(\sum_{z'\in\mathcal{N}_z} a_{z'\to z} \boldsymbol{h}_{z'}^{k-1} \right) - \phi(\mathbf{0}) \right\|_2 \\
&\leq C_\phi \cdot \sum_{z\in\mathcal{Z}} \left\| \sum_{z'\in\mathcal{N}_z} a_{z'\to z} \boldsymbol{h}_{z'}^{k-1} \right\|_2 \\
&\leq \gamma C_\phi \cdot \sum_{z\in\mathcal{Z}} \sum_{z'\in\mathcal{N}_z} \left\|\boldsymbol{h}_{z'}^{k-1}\right\|_2 \\
&\leq \gamma C_\phi d_{\max} \cdot \sum_{z\in\mathcal{Z}} \left\|\boldsymbol{h}_z^{k-1}\right\|_2 \\
&\leq (\gamma C_\phi d_{\max})^k \cdot \sum_{z\in\mathcal{Z}} \left\|\boldsymbol{h}_z^0\right\|_2 \\
&\leq (\gamma C_\phi d_{\max})^k \cdot NTB_X . \qquad (18)
\end{aligned}$$

Therefore, expanding the recursion in Eq. 16, we have

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{M}}\left[\sum_{z\in\mathcal{Z}_o^k}\left\|\tilde{\boldsymbol{h}}_z^k-\boldsymbol{h}_z^k\right\|_2\right]\\
&\leq \mathbb{E}_{\boldsymbol{M}}\left[\gamma C_\phi d_{\max}\cdot\sum_{z\in\mathcal{Z}_o^{k-1}}\left\|\tilde{\boldsymbol{h}}_{z'}^{k-1}-\boldsymbol{h}_{z'}^{k-1}\right\|_2+\gamma C_\phi\cdot\sum_{z\in\mathcal{Z}_o^k}\left(\sum_{z'\in\mathcal{N}_z\setminus\tilde{\mathcal{N}}_z^k}\left\|\boldsymbol{h}_{z'}^{k-1}\right\|_2+\sum_{z'\in\tilde{\mathcal{N}}_z}\eta\cdot\left\|\boldsymbol{h}_{z'}^{k-1}\right\|_2\right)\right]\\
&\leq \gamma C_\phi d_{\max}\cdot\left(\mathbb{E}_{\boldsymbol{M}}\left[\sum_{z\in\mathcal{Z}_o^{k-1}}\left\|\tilde{\boldsymbol{h}}_{z'}^{k-1}-\boldsymbol{h}_{z'}^{k-1}\right\|_2\right]+\left[\rho^{\tau_{\min}^{k-1}}+\eta\cdot(1-\rho^{\tau_{\max}^{k-1}})\right]\cdot\sum_{z\in\mathcal{Z}}\left\|\boldsymbol{h}_z^{k-1}\right\|_2\right)\\
&\leq \gamma C_\phi d_{\max}\cdot\mathbb{E}_{\boldsymbol{M}}\left[\sum_{z\in\mathcal{Z}_o^{k-1}}\left\|\tilde{\boldsymbol{h}}_{z'}^{k-1}-\boldsymbol{h}_{z'}^{k-1}\right\|_2\right]+\left[\rho^{\tau_{\min}^{k-1}}+\eta\cdot(1-\rho^{\tau_{\max}^{k-1}})\right]\cdot(\gamma C_\phi d_{\max})^k\cdot NTB_X.
\end{aligned}
\tag{19}
$$

Then we can derive

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{M}}\left[\sum_{z\in\mathcal{Z}}\left\|\tilde{\boldsymbol{h}}_z^K-\boldsymbol{h}_z^k\right\|_2\right]=\mathbb{E}_{\boldsymbol{M}}\left[\sum_{z\in\mathcal{Z}_o^K}\left\|\tilde{\boldsymbol{h}}_z^K-\boldsymbol{h}_z^k\right\|_2\right]\\
&\leq \gamma C_\phi d_{\max}\cdot\mathbb{E}_{\boldsymbol{M}}\left[\sum_{z\in\mathcal{Z}_o^{k-1}}\left\|\tilde{\boldsymbol{h}}_{z'}^{k-1}-\boldsymbol{h}_{z'}^{k-1}\right\|_2\right]+\left[\rho^{\tau_{\min}^{k-1}}+\eta\cdot(1-\rho^{\tau_{\max}^{k-1}})\right]\cdot(\gamma C_\phi d_{\max})^k\cdot NTB_X\\
&\leq (\gamma C_\phi d_{\max})^K\cdot NTB_X\cdot\sum_{k=1}^{K}\left[\rho^{\tau_{\min}^{k-1}}+\eta\cdot(1-\rho^{\tau_{\max}^{k-1}})\right].
\end{aligned}
\tag{20}
$$

Substituting Eq.20 into Eq.12, we have

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{M}}\left|L(\tilde{\boldsymbol{Y}},\boldsymbol{Y})-L(\hat{\boldsymbol{Y}},\boldsymbol{Y})\right|&\leq C_l\cdot C_\phi\cdot B_d\cdot\frac{1}{NT}\cdot\mathbb{E}_{\boldsymbol{M}}\left[\sum_{z\in\mathcal{Z}}\left\|\tilde{\boldsymbol{h}}_z^K-\boldsymbol{h}_z^K\right\|_2\right]\\
&\leq C_l\cdot C_\phi\cdot B_d\cdot\frac{1}{NT}\cdot(\gamma C_\phi d_{\max})^K\cdot NTB_X\cdot\sum_{k=1}^{K}\left[\rho^{\tau_{\min}^{k-1}}+\eta\cdot(1-\rho^{\tau_{\max}^{k-1}})\right]\\
&\leq C_l\cdot\gamma^K\cdot C_\phi^{K+1}\cdot(d_{\max})^K\cdot B_XB_d\cdot\sum_{k=1}^{K}\left[\rho^{\tau_{\min}^{k-1}}+\eta\cdot(1-\rho^{\tau_{\max}^{k-1}})\right].
\end{aligned}
$$

Let $\mathcal{C}=C_l\cdot\gamma^K\cdot C_\phi^{K+1}\cdot(d_{\max})^K\cdot B_XB_d$, we finally derive the follows inequality.

$$
\mathbb{E}_{\boldsymbol{M}}\left|L(\tilde{\boldsymbol{Y}},\boldsymbol{Y})-L(\hat{\boldsymbol{Y}},\boldsymbol{Y})\right|\leq\mathcal{C}\cdot\sum_{k=1}^{K}\left[\rho^{\tau_{\min}^{k-1}}+\eta\cdot(1-\rho^{\tau_{\max}^{k-1}})\right].
$$

□

For the second term in Eq.8, $R_{\mathcal{D}}(f)-\hat{R}_S(f)$ represents the generalization performance of $f\in\mathcal{F}$ over complete dataset. Model complexity measurements, such as VC-Dimension [27], Rademacher complexity [28] and Covering number [29] are related to the generalization ability. Therefore, they can provide useful information about number of training samples m required for PAC-learnability under complete data. Rademacher complexity is frequently applied to investigate the generalization performances of GNN-based models. The Rademacher complexity-based generalization bound as given in [15] is:

Lemma B.3. *Let $\mathcal{H}$ be a function class mapping from $\mathcal{X}$ to $[0,1]$. Then, for any $\delta>0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size m, for any $h\in\mathcal{H}$, the following*

https://doi.org/10.52202/079017-2982

inequality holds.

$$R_{\mathcal{D}}(h) \leq \hat{R}_S(h) + 2\Re_m(\mathcal{H}) + \sqrt{\frac{\log 1\backslash\delta}{2m}}$$

As our assumptions, the loss function $l : \mathbb{R} \times \mathcal{Y} \to [0,1]$ is C_l-lipschitz continuous. Then we can derive the following corollary, which show that generalization gap is small when provided large enough training samples.

Corollary B.4. *Let $\mathcal{F} : \mathcal{X} \to [0,1]$ be a function class for spatio-temporal data imputation. function class mapping from $\mathcal{X}$ to $[0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size m, for any $h \in \mathcal{H}$, the following inequality holds.*

$$P\left(\left|R_{\mathcal{D}}(f) - \hat{R}_S(f)\right| \geq \epsilon\right) \leq e^{-2m(\epsilon - 2C_l\Re_m(\mathcal{F}))}.$$

Proof. Using Lemma B.3 and contraction lemma[15], we have

$$\begin{aligned} R_{\mathcal{D}}(f) &\leq \hat{R}_S(f) + 2\Re_m(l \circ \mathcal{F}) + \sqrt{\frac{\log 1\backslash\delta}{2m}} \\ &\leq \hat{R}_S(f) + 2C_l\Re_m(\mathcal{F}) + \sqrt{\frac{\log 1\backslash\delta}{2m}}. \end{aligned}$$

Therefore, we can rewrite this inequality then we have for any $f \in \mathcal{F}$

$$P\left(\left|R_{\mathcal{D}}(f) - \hat{R}_S(f)\right| \geq \epsilon\right) \leq e^{-2m(\epsilon - 2C_l\Re_m(\mathcal{F}))}.$$

□

Using Corollary B.2 and B.4, we have

$$\begin{aligned} &P\left(\left|R_{\tilde{\mathcal{D}}}(f) - \hat{R}_{\tilde{S}}(f)\right| \geq \epsilon\right) \\ &\leq P\left(\left|R_{\mathcal{D}}(f) - \hat{R}_S(f)\right| \geq \epsilon - \left|R_{\tilde{\mathcal{D}}}(f) - R_{\mathcal{D}}(f) + \hat{R}_S(f) - \hat{R}_{\tilde{S}}(f)\right|\right) \\ &\leq P\left(\left|R_{\mathcal{D}}(f) - \hat{R}_S(f)\right| \geq \epsilon - 2\mathcal{C}\cdot\sum_{k=1}^{K}\left[\rho^{\tau_{\min}^{k-1}} + \eta\cdot(1-\rho^{\tau_{\max}^{k-1}})\right]\right) \\ &\leq \exp\left(-2m\left[\epsilon - 2\mathcal{C}\cdot\sum_{k=1}^{K}\left(\eta - \eta\rho^{\tau_{\max}^{k-1}} + \rho^{\tau_{\min}^{k-1}}\right) - 2C_l\Re_m(\mathcal{F})\right]\right). \end{aligned}$$

Then we derive the number of required samples m:

$$\begin{aligned} &\exp\left(-2m\left[\epsilon - 2\mathcal{C}\cdot\sum_{k=1}^{K}\left(\eta - \eta\rho^{\tau_{\max}^{k-1}} + \rho^{\tau_{\min}^{k-1}}\right) - 2C_l\Re_m(\mathcal{F})\right]\right) \leq \delta \\ &\Rightarrow \exp\left(2m\left[\epsilon - 2\mathcal{C}\cdot\sum_{k=1}^{K}\left(\eta - \eta\rho^{\tau_{\max}^{k-1}} + \rho^{\tau_{\min}^{k-1}}\right) - 2C_l\Re_m(\mathcal{F})\right]\right) \geq \frac{1}{\delta} \\ &\Rightarrow 2m\left[\epsilon - 2\mathcal{C}\cdot\sum_{k=1}^{K}\left(\eta - \eta\rho^{\tau_{\max}^{k-1}} + \rho^{\tau_{\min}^{k-1}}\right) - 2C_l\Re_m(\mathcal{F})\right] \geq \log(\frac{1}{\delta}) \\ &\Rightarrow m \geq \frac{1}{2\epsilon - 4\mathcal{C}\cdot\sum_{k=1}^{K}\left(\eta - \eta\rho^{\tau_{\max}^{k-1}} + \rho^{\tau_{\min}^{k-1}}\right) - 4C_l\Re_m(\mathcal{F})}\cdot\log\sqrt{\frac{1}{\delta}} \\ &\Rightarrow m \geq \frac{1}{2\epsilon - 4\eta\mathcal{C}K - 4(1-\eta)\mathcal{C}\cdot\sum_{k=1}^{K}\rho^{\tau_{\max}^{k-1}} - 4C_l\Re_m(\mathcal{F})}\cdot\log\sqrt{\frac{1}{\delta}} \\ &\Rightarrow m \geq \frac{1}{2\epsilon - 4\eta\mathcal{C}K - 4(1-\eta)\mathcal{C}K\cdot\rho^{\tau_{\max}^{K-1}} - 4C_l\Re_m(\mathcal{F})}\cdot\log\sqrt{\frac{1}{\delta}} \end{aligned}$$

For convenience, we let $\tau = \tau_{\max}^{K-1}$.

C Proof of Proposition 5.1

Proof. Similar to proof of Proposition 5.1, we first derive the following upper bound for the difference between model performances under complete and sparse data. For any $\boldsymbol{M} \in \mathcal{M}$, we have

$$
\begin{aligned}
&\frac{1}{NT} \cdot \sum_{z \in \mathcal{Z}} \left\| \tilde{\boldsymbol{h}}_z - \boldsymbol{h}_z \right\|_2 \\
&= \frac{1}{NT} \cdot \sum_{z \in \mathcal{Z}} \left\| \sigma(\tilde{\boldsymbol{a}}_z) \boldsymbol{V} - \sigma(\boldsymbol{a}_z) \boldsymbol{V} \right\|_2 \\
&\leq \frac{1}{NT} \cdot \sum_{z \in \mathcal{Z}} \left\| \sum_{z' \in \mathcal{Z}_o} \frac{a_{z' \to z}}{\sum_{z'' \in \mathcal{Z}_o} a_{z'' \to z}} \boldsymbol{v}_{z'} - \sum_{z' \in \mathcal{Z}} \frac{a_{z' \to z}}{\sum_{z'' \in \mathcal{Z}} a_{z'' \to z}} \boldsymbol{v}_{z'} \right\|_2 \\
&\leq \frac{B_v}{NT} \cdot \sum_{z \in \mathcal{Z}} \left| \sum_{z' \in \mathcal{Z}_o} \left(\frac{a_{z' \to z}}{\sum_{z'' \in \mathcal{Z}_o} a_{z'' \to z}} - \frac{a_{z' \to z}}{\sum_{z'' \in \mathcal{Z}} a_{z'' \to z}} \right) \right| + \frac{1}{NT} \cdot \sum_{z \in \mathcal{Z}} \left| \sum_{z' \in \mathcal{Z} \setminus \mathcal{Z}_o} \frac{a_{z' \to z}}{\sum_{z'' \in \mathcal{Z}} a_{z'' \to z}} \right|,
\end{aligned}
$$

where $\max_{z \in \mathcal{Z}} \|\boldsymbol{v}_z\|_2 \leq B_v$. Then we have

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{M}} \left[\frac{1}{NT} \cdot \sum_{z \in \mathcal{Z}} \left\| \tilde{\boldsymbol{h}}_z - \boldsymbol{h}_z \right\|_2 \right] \\
&\leq \frac{B_v}{NT} \cdot \sum_{z \in \mathcal{Z}} \mathbb{E}_{\boldsymbol{M}} \left[\left| \sum_{z' \in \mathcal{Z}_o} \frac{a_{z' \to z} \cdot \left(\sum_{z'' \in \mathcal{Z} \setminus \mathcal{Z}_o} a_{z'' \to z} \right)}{\left(\sum_{z'' \in \mathcal{Z}_o} a_{z'' \to z} \right) \cdot \left(\sum_{z'' \in \mathcal{Z}} a_{z'' \to z} \right)} \right| + \left| \sum_{z' \in \mathcal{Z} \setminus \mathcal{Z}_o} \frac{a_{z' \to z}}{\sum_{z'' \in \mathcal{Z}} a_{z'' \to z}} \right| \right] \\
&= \frac{2B_v}{NT} \cdot \sum_{z \in \mathcal{Z}} \mathbb{E}_{\boldsymbol{M}} \left| \frac{\sum_{z' \in \mathcal{Z} \setminus \mathcal{Z}_o} a_{z' \to z}}{\sum_{z' \in \mathcal{Z}} a_{z' \to z}} \right| \\
&\leq 2B_v \cdot \rho
\end{aligned}
$$

Similar to proof of Proposition 4.3,

$$
\begin{aligned}
&P\left(\left| R_{\tilde{\mathcal{D}}}(f) - \hat{R}_{\tilde{S}}(f) \right| \geq \epsilon \right) \leq \exp\left(-2m \left[\epsilon - 4C_l \cdot C_\phi \cdot B_d \cdot B_v \cdot \rho - 2C_l \Re_m(\mathcal{F}) \right] \right) \\
&\Rightarrow P\left(\left| R_{\tilde{\mathcal{D}}}(f) - \hat{R}_{\tilde{S}}(f) \right| \geq \epsilon \right) \leq \exp\left(-2m \left[\epsilon - 4\mathcal{C} \cdot \rho - 2C_l \Re_m(\mathcal{F}) \right] \right) \\
&\Rightarrow \exp\left(-2m \left[\epsilon - 4\mathcal{C} \cdot \rho - 2C_l \Re_m(\mathcal{F}) \right] \right) \leq \delta \\
&\Rightarrow 2m \left[\epsilon - 4\mathcal{C} \cdot \rho - 2C_l \Re_m(\mathcal{F}) \right] \geq \log \frac{1}{\delta} \\
&\Rightarrow m \geq \frac{\log 1/\delta}{2 \left[\epsilon - 4\mathcal{C} \cdot \rho - 2C_l \Re_m(\mathcal{F}) \right]},
\end{aligned}
$$

where $\mathcal{C} = C_l \cdot C_\phi \cdot B_d \cdot B_v$. □

https://doi.org/10.52202/079017-2982

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We summarize the motivation and contribution of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification:

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Appendix

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Reproducible code and Experimental settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-2982

Answer: [Yes]

Justification: Reproducible code and Experimental settings.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Reproducible code and Experimental settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We run all the experiments with 5 different random seeds and present the results with mean ± standard deviation (std).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experimental settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We point out that exploring spatially missing data can lower the barrier of some realistic applications, such as intelligent transportation.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/079017-2982

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have illustrated that some experiments on the traffic dataset follow SPIN.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2982