Dual-stream Spatio-temporal Contrastive Learning for Long-term Traffic Anomaly Detection

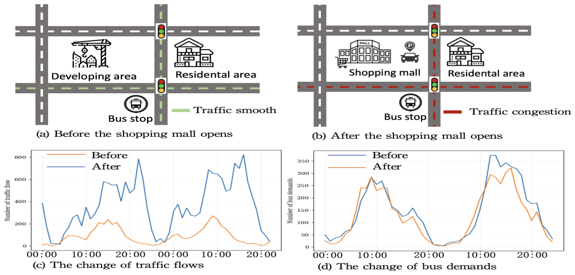
Anonymous submission

Abstract

Detection of traffic anomalies has attracted widespread interest in recent years. Among them, long-term recurrent traffic anomaly is valuable since it aims to identify mismatched traffic patterns from expected road/demand service, which could help in many urban planning applications. This task is challenging due to several factors: (i) the complexity of long-range spatio-temporal interactions, (ii) the implicit incompatibility from corresponding service levels, and (iii) the disturbances from non-recurrent and random traffic abnormals. Hence, we propose a dual-stream spatio-temporal contrastive learning method (DCSL) to detect long-term traffic abnormalities considering both their distances to the global context and expected traffic capabilities, with handled patched embedding data. We propose a dynamic stream to learn the complex and long spatial-temporal dependencies based on a novel attention block and a contrastive soft clustering self-supervised algorithm to extract the spatial heterogeneity adaptively. Further, we propose a static stream to learn corresponding relationships with the respective service by a novel scalable multi-view graph convolutional network. Finally, we design a novel detector consisting of two spatio-temporal contrastive losses and a reconstruction loss to effectively detect long-term traffic anomalies. Extensive experiment results on real-world datasets demonstrate that our method has significant advantages over the existing models.

Introduction

In recent years, a growing number of urban traffic anomalies have contributed to an increase in traffic congestion and accidents, affecting transportation efficiency and public safety. Therefore, traffic anomaly detection plays a vital role in intelligent transportation management, such as adjusting signal timings and rerouting traffic in advance. Urban traffic anomalies can be divided into two categories: (i) short-term anomalies, which are usually caused by unexpected events, such as traffic accidents and special social events, and (ii) long-term anomalies, such as recurrent traffic congestion, which is a long-term mismatch to expected traffic service due to the traffic structural changes and development of urban layouts. Reports of Chow et al. (2014) highlighted that 85% of the congestion and high-risk traffic accidents on the urban road network are long-term and recurrent, and occur in specific locations. For example, as illustrated in Figure 1, an area where a new shopping mall is built experiences a dramatic increase in traffic demand. However, the existing public transit routes have yet to be modified to accommodate this change, and the traffic signal system at intersections remains ill-suited to the current demands that lead to unsatisfied bus service and long-term recurrent traffic congestion. Therefore, the accurate detection of long-term traffic anomalies can significantly enhance traffic fluidity and road utilization efficiency, and fundamentally reduce traffic congestion and anomalies.



(a) Before the shopping mall opens

(b) After the shopping mall opens

(c) The change of traffic flows

(d) The change of bus demands

Figure 1: An example of the long-term traffic anomaly.

Various data-driven and time series anomaly detection methods have been developed in the past for short-term non-recurrent traffic anomalies, such as traffic congestion and accidents. However, these methods are difficult to use in long-term anomaly detection due to the following issues.

First, identification of long-range spatio-temporal discrepancy across regions. Compared to short-term non-recurrent traffic anomaly detection that focuses on sudden temporal fluctuations, long-term traffic anomalies are more challenging since they have relatively stable patterns at temporal scales and are also influenced by complex spatial region associations, affecting population migration, urban planning, weather, etc. For example, similar traffic dynamics that occurred in a residential area and a park could be represented as normal or abnormal signals because of the corresponding functionalities. Therefore, the traffic recurrent anomaly detection problem emphasizes extracting spatio-temporal discrepancy from global traffic neighbors. It requires modeling the long-range spatio-temporal interactions and distancing them from the anomalous regions and other counterparts while avoiding disturbances caused by non-recurrent traffic accidents.

Second, the discovery of incompatibility between expected and actual traffic flows from heterogeneous views. As illustrated, the long-term traffic outliers are usually recurrent and observed as mismatches to the corresponding level of service on the road segments. The patterns of long-term traffic anomalies are diverse since they are caused by different types of changed urban context information, such as imbalanced bus serviceability or sudden increased population of regions. On the one hand, the model requires learning the static serviceability of regions/road segments based on multiple structure views of factors; on the other hand, it needs to learn the dynamic data-driven spatial-temporal heterogeneity of regions/road segments based on adaptive and self-supervised learning. Then, how to detect the contrasting differences between static association and dynamic self-learned spatio-temporal dependencies is a unique challenge.

To address the aforementioned challenges, we propose a Dual-stream Spatio-temporal Contrastive Learning method (DSCL) to detect abnormal long-term traffic flows from normal ones based on both their distances to the global context and expected traffic service capabilities. A dynamic global-context stream is proposed to learn the complex and long spatial and temporal interactions via a spatial-temporal self-attention mechanism. To capture the latent data-driven spatial heterogeneity among regions, we propose a self-supervised contrastive soft clustering algorithm to make it aware of the diverse spatial patterns among different regions. Then, we propose static service-flow steam to learn the corresponding relationships between traffic flow and its expected service via a novel scalable multi-view graph convolutional networks and a temporal attention mechanism. In the detection part, we propose a joint detector that consists of 3 losses. A reconstruction loss is used to identify traffic anomalies from the fusing representation of the static and dynamic streams. To achieve the contrastive goal between dynamic abnormal traffic flows and corresponding regional traffic services, we design two contrastive losses to enhance the ability of long-term traffic anomaly detection.

- We design a contrastive learning-based dual-stream method to detect long-term traffic anomalies from dynamic and static views. This can effectively capture the diverse abnormal traffic flows from normal traffic flow distribution and corresponding location service.
- In the dynamic stream, we propose a heterogeneity-aware contrastive soft clustering algorithm and attention blocks to effectively extract complex long-range dependencies from global spatio-temporal contexts and filter out noise simultaneously. In the static stream, we present a scalable GCN and a temporal attention mechanism to model corresponding associations of traffic flows and their expected service of regions from multiple views.
- We proposed an anomaly detector to qualify the mismatch between dynamic and static streams, where we adaptively combine two effective contrastive losses and a reconstruction loss as the criterion of long-term anomaly.
- We evaluate our proposed method on real-world datasets. Extensive comparative experiments demonstrate the superiority of our method over state-of-the-art baselines.

Related Works

Traffic Anomaly Detection

Traffic anomaly detection, as an important task in Urban planning, has attracted much attention in recent years. Regarding the categories of traffic anomalies, there are two common types: short-term nonrecurrent and long-term recurrent traffic anomalies. Existing traffic anomaly detection methods mostly focus on discovering short-term anomalies. Generally, current traffic anomaly detection works can be divided into three categories: statistical, matrix-based, and spatio-temporal feature-based methods (Zhang et al. 2020).

Statistical models were commonly used to detect traffic outliers in the early years, such as the hidden Markov model (Witayangkurn et al. 2013) and ARIMA (Mihaita, Li, and Rizoiu 2020). In recent years, matrix-based methods have regarded traffic flows as a combination of some pattern distributions of time series, usually used to model the complex distribution of abnormal traffic flows in different regions and time slots. For example, (Lin et al. 2018) applied a non-negative decomposition to derive latent mobility patterns, and (Wang et al. 2019) presented a context-aware Tucker factorization method to extract interpretable urban dynamics based on multiple neighboring relations. Spatio-temporal feature-based methods detect traffic anomalies by extracting abnormal complex spatial and temporal correlations from traffic data. For example, Zhang, Zheng, and Yu proposed a similarity-based algorithm to estimate traffic anomaly scores via a one-class SVM. Deng et al. proposed an end-to-end anomaly detection discriminator by learning a flexible anomaly score based on spatio-temporal features.

Contrastive Representation Learning

Contrastive representation learning aims to design objective functions in latent space by contrasting positive and negative samples (Caron et al. 2020; Chen et al. 2020; He et al. 2020). Since it is unfeasible to access out-of-distribution (OOD) data in most real-world scenarios, the ability of contrastive self-supervised learning has received attention from anomaly detection lately (Cho, Seol, and goo Lee 2023). For example, Liu et al. detects anomalies on attributed networks by proposing a contrastive self-supervised learning framework to discriminate the agreement between nodes and subgraphs. Yang et al. presented a dual-branch structure to learn a permutation invariant representation between normal points and anomalies.

However, these existing methods either need to be revised in modeling long-range traffic dynamics or consider the interrelationships with the expected traffic service of regions. They lack the ability to detect long-term traffic anomalies. In our work, we extract long-term urban context information from multiple views based on proposed novel spatio-temporal attention and contrastive self-learning representations. We also design a flexible anomaly score that not only considers the reconstruction loss from the global context but also proposes contrastive loss to distinguish the abnormal traffic service performance.

Preliminaries

We begin with some useful definitions and notations.

Definition 1 (Traffic network) *We define traffic network as a weighted multi-view graph $\mathcal{G} = \{\mathcal{V}, (\mathcal{E}_i, \mathbf{A}_i) | i = 1, 2, \cdots, N_{view}\}$ to describe the comprehensive structure of a transportation system, where $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$ is a set of nodes that represents sensors in a transport network with the size of N, and different views are used to describe the nodal heterogeneous properties, where $\mathcal{E}_i$ and $\mathbf{A}_i \in \mathbb{R}^{N \times N}$ denote the set of edges and the adjacency matrix in the i-th view, respectively.*

Definition 2 (Traffic signal) *Traffic signal is defined as a tensor $\mathcal{X} \in \mathbb{R}^{N \times T \times D}$ representing citywide D-dimensional traffic signal on N urban regions over T continuous time steps. We denote the traffic signal information of nodes at the t-th time slot as $\mathbf{X}_t \in \mathbb{R}^{N \times D}$ and the traffic signal information of all times slots at the n-th node as $\mathbf{X}^{(n)} \in \mathbb{R}^{T \times D}$.*

Definition 3 (Long-term traffic anomaly) *Given a region or sensor, the long-term anomaly measures the incompatibility between its real traffic signals and expected/corresponding service levels from a long-term recurrent perspective.*

Definition 4 (Long-term traffic anomaly detection) *Given a traffic graph $\mathcal{G}$ and its corresponding traffic signals $\mathcal{X}$, we aim to learn an anomaly score function $Score(\cdot)$. the nodes in $\{v \in \mathcal{V} | Score(v) \geq \epsilon\}$ are regions or sensors with long-term traffic anomalies.*

It is worth noting that short-term traffic anomaly detection only considers this incompatibility on a single time step.

Methodology

To better identify the discrepancies between its actual traffic signals and the expected corresponding service levels from a long-term perspective, we propose DSCL, as shown in Figure 2, which consists of a patching spatio-temporal embedding layer, a dynamic global-context stream component, a static flow-service stream component, and a long-term traffic anomaly detector. We will describe these modules in detail.

Patching Spatio-temporal Embedding

As illustrated in Figure 1, recurrent traffic anomalies usually occur with the development of cities, due to many reasons, such as the migration of citizens, changing of region functionalities, and improper traffic signals. To discover these traffic dynamics, the detection model requires the extraction of stable and long-range spatio-temporal dependencies. Therefore, an ideal urban traffic dynamic representation should be able to reduce the disturbance caused by short-term anomaly noise and effectively learn the dependencies of time and space.

To realize this, we propose a convolutional patching spatio-temporal embedding method, which consists of a temporal causal convolution and a patching mechanism. The introduction of causal convolution can improve the awareness of temporal self-attention to local context and compute their similarities by their local context information, e.g., local shapes, instead of point-wise values, making the model robust to short-term anomalies (Li et al. 2019).

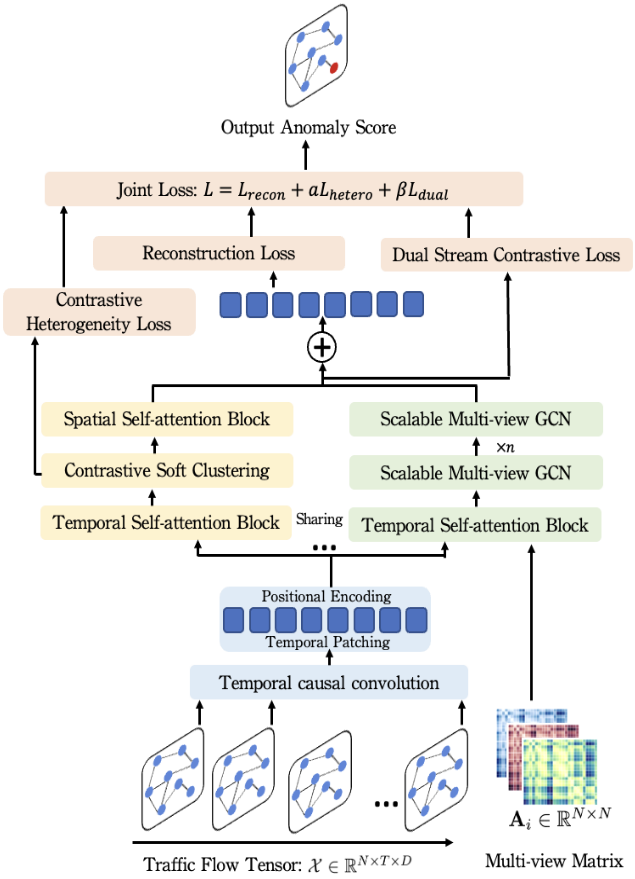

Output Anomaly Score
Joint Loss: $L = L_{recon} + aL_{hetero} + \beta L_{dual}$
Reconstruction Loss
Dual Stream Contrastive Loss
Contrastive Heterogeneity Loss
Spatial Self-attention Block
Scalable Multi-view GCN
×n
Contrastive Soft Clustering
Scalable Multi-view GCN
Temporal Self-attention Block
Sharing
Temporal Self-attention Block
Positional Encoding
Temporal Patching
Temporal causal convolution
$\mathbf{A}_i \in \mathbb{R}^{N \times N}$
Traffic Flow Tensor: $\mathcal{X} \in \mathbb{R}^{N \times T \times D}$
Multi-view Matrix

Figure 2: The framework of DSCL. DSCL has four components. The blue part is the patching spatio-temporal embedding layers, the yellow part is the dynamic global-context stream component, the green part is the static flow-service stream component, and the orange part is the long-term traffic anomaly detector.

The patching mechanism partitions traffic signal data along the temporal axis into non-overlapping patches (Nie et al. 2023). Denoting the patch length as P, the number of input tokens reduces from T to T/P, which implies the memory usage and computational complexity of the latter temporal attention map are quadratically reduced by a factor of P. We follow the channel-independence design, i.e., each input token only contains information from a single channel. We consider the single-channel case for simplicity. Formally, given the patch length P, after causal convolution and patching, for each channel, the input signal is transformed into a set of patches $\mathcal{P} \in \mathbb{R}^{N \times N_P \times P}$, where $N_P = \lceil T/P \rceil$ is the number of patches. The patches are mapped to the latent space of dimension d via a trainable linear projection $\mathbf{W}_p \in \mathbb{R}^{P \times d}$, and a learnable addictive positional encoding $\mathbf{W}_{\text{pos}} \in \mathbb{R}^{N \times N_P \times d}$ is applied to monitor the spatial and temporal order of patches, which can be written as:

$$\hat{\mathcal{P}} = \mathcal{P}\mathbf{W}_p + \mathbf{W}_{\text{pos}} \in \mathbb{R}^{N \times N_P \times d} .$$

Dynamic Global-context Stream

The dynamic global context stream is proposed to learn the global and long-term spatio-temporal interactions among regions that consists of spatial and temporal self-attention blocks, with a contrastive soft clustering layer inserted in, which is different from existing sequential spatial-temporal Transformer structures for prediction tasks (Xu et al. 2020), in order to learn the urban regional functionalities and semantics adaptively. Here, we call it *regional heterogeneity*.

Temporal Self-Attention Block Given the embedding of n-th node, $\hat{\mathcal{P}}^{(n)} \in \mathbb{R}^{N_P \times d}$, from the patching spatio-temporal embedding layer, we feed it into a temporal self-attention encoder. Each head in the multi-head attention will transform it into a query matrix $\mathbf{Q}^{(n)} = \hat{\mathcal{P}}^{(n)}\mathbf{W}^Q$, a key matrix $\mathbf{K}^{(n)} = \hat{\mathcal{P}}^{(n)}\mathbf{W}^K$ and a value matrix $\mathbf{V}^{(n)} = \hat{\mathcal{P}}^{(n)}\mathbf{W}^V$, where $\mathbf{W}^Q, \mathbf{W}^K \in \mathbb{R}^{d \times d_k}$ and $\mathbf{W}^V \in \mathbb{R}^{d \times d}$. After that, a scaled dot-product is performed to obtain the attention output $\mathbf{O}^{(n)} \in \mathbb{R}^{N_P \times D}$ as follows:

$$\mathbf{O}^{(n)} = \text{Softmax}\left(\frac{\mathbf{Q}^{(n)}\mathbf{K}^{(n)\top}}{\sqrt{d_k}}\right)\mathbf{V}^{(n)} \ .$$

The temporal multi-head attention block also includes a LayerNorm and a feed-forward network with residual connections. We denote its output for the n-th node as $\mathbf{Z}^{(n)}$ and let the concatenation $\mathcal{Z} = [\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \cdots, \mathbf{Z}^{(N)}]$ denote the representation of overall graph nodes.

Contrastive soft clustering A key observation is that different regions have quite different traffic flow distributions due to their functionalities. Such spatial heterogeneity significantly influences the representation quality of traffic signals. To effectively extract the abnormal spatio-temporal patterns in a traffic network, we design an adaptive soft clustering-based self-supervised learning task, which enables the temporal region embeddings to be semantic-aware with auxiliary self-supervised signals before embedding the input into spatial self-attention. Specifically, we map them into multiple latent representation spaces corresponding to diverse urban region functionalities.

We generate K cluster centers $\mathcal{C} = \{\mathbf{C}_1, \cdots, \mathbf{C}_K\}$ for the region clusters, where $\mathbf{C}_k \in \mathbb{R}^{T \times d}$ for $k = 1, 2, \cdots, K$. The cluster assignment for all nodes is defined as follows:

$$\mathbf{S} = \left[\langle \mathbf{Z}^{(i)}, \mathbf{C}_j \rangle_F\right]_{N \times K} \ ,$$

where $\langle \mathbf{A}, \mathbf{B} \rangle_F = \sum_{i,j} \mathbf{A}(i,j)\mathbf{B}(i,j)$, which can be viewed as the inner-product of two bi-dimensional vectors. To extract the long-term traffic anomaly by understanding the spatial-temporal heterogeneity of regions, we propose an auxiliary learning optimization process for this clustering paradigm as an additional contrastive heterogeneity loss $\mathcal{L}_{\text{hetero}}$. The optimization process is formulated as follows:

$$\mathcal{L}_{\text{hetero}} = -\sum_{n=1}^{N}\sum_{k=1}^{K} \mathbf{S}(n,k) \log \frac{\exp(\mathbf{S}(n,k)/\gamma)}{\sum_{j=1}^{K} \exp(\mathbf{S}(n,j)/\gamma)} \ , \tag{1}$$

where γ is the temperature parameter to control the smoothing degree of the softmax output.

Spatial Self-Attention Block Denote the representation of the t-th patch for all nodes as $\mathcal{Z}_t \in \mathbb{R}^{N \times d}$, it is then presented to the spatial self-attention encoder. Each head in the multi-head attention will transform it into a query matrix $\mathbf{Q}_t = \mathcal{Z}_t \bar{\mathbf{W}}_h^Q$, a key matrix $\mathbf{K}_t = \mathcal{Z}_t \bar{\mathbf{W}}^K$ and a value matrix $\mathbf{V}_t = \mathcal{Z}_t \bar{\mathbf{W}}^V$, where $\bar{\mathbf{W}}^Q, \bar{\mathbf{W}}^K \in \mathbb{R}^{d \times d_k}$ and $\bar{\mathbf{W}}^V \in \mathbb{R}^{d \times d}$. After that, a scaled dot-prod operation is performed to obtain the attention output $\mathbf{O}_t \in \mathbb{R}^{N \times D}$:

$$\mathbf{O}_t = \mathbf{Attn}_t \mathbf{V}_t = \text{Softmax}\left(\frac{\mathbf{Q}_t \mathbf{K}_t^\top}{\sqrt{d_k}}\right)\mathbf{V}_t \ .$$

Similar to the temporal self-attention, after passing through a LayerNorm layer and a feed-forward network with residual connections, we obtain the representation of the t-th patch, denoted as $\bar{\mathbf{Z}}_t$, and the concatenation $\bar{\mathcal{Z}} = [\bar{\mathbf{Z}}_1, \bar{\mathbf{Z}}_2, \cdots, \bar{\mathbf{Z}}_{N_P}]$, which denotes the representation of overall patches. Finally, the matrix $\mathcal{A}_D = \sum_{t=1}^{N_P} \mathbf{Attn}_t / N_P$ can be used to characterize the global dynamic spatio-temporal dependencies. In the multi-head case, we take it as the element-wise mean of $\mathcal{A}_D$ of each head.

Static Service-flow Stream

The static service-flow stream aims to learn the corresponding relationships between traffic flow and its expected service level, which relies on the prior regional relationships based on distance, connectivity, and functionality similarity. To capture these spatial dependencies comprehensively, in this paper, we propose a scalable multi-view graph convolutional network to learn the static spatial dependency and the urban expected service level with diverse heterogeneous relationships. Specifically, given the adjacency matrices $\mathbf{A}_k, k = 1, \cdots, N_{\text{view}}$, we first use a learnable Gaussian kernel to calculate the region-adaptive spatial relationship:

$$\hat{\mathbf{A}}_k = \text{Scale}\left(\exp\left(-\frac{\mathbf{A}_k(i,j)^2}{2\boldsymbol{\sigma}(k,i)^2}\right)\right)_{N \times N} \ ,$$

where $\boldsymbol{\sigma} \in \mathbb{R}_{++}^{N_{\text{view}} \times N}$ is the learnable parameter to reflect the relational sensitivity to other nodes on different views. We use $\text{Scale}(\cdot)$ to transform the static dependency to discrete distributions $\hat{\mathbf{A}}_k$ by dividing the row sum.

We then exploit the scalable graph structure to learn the traffic flow with the expected service level using graph convolution. After obtaining the embedding $\mathcal{Z}_t \in \mathbb{R}^{N \times d}$ of the t-th patch of all nodes from a temporal self-attention layer, which shares the same parameters within the dynamic global-context stream, for the k-th view, we conduct it as:

$$\hat{\mathbf{Z}}_t = \hat{\mathbf{A}}_k f(\hat{\mathbf{A}}_k \mathcal{Z}_t \mathbf{W}_k^{(1)})\mathbf{W}_k^{(2)} \ ,$$

where $\mathbf{W}_k^{(1)}, \mathbf{W}_k^{(2)}$ are weight matrices for graph convolution on the k-th view, and f is the activation function. The concatenation $\hat{\mathcal{Z}} = [\hat{\mathbf{Z}}_1, \cdots, \hat{\mathbf{Z}}_{N_P}]$ denotes the representation of overall patches.

We then fuse the scalable multi-view graph matrix with learnable parameters to describe the static regional spatio-temporal dependency $\mathcal{A}_S$ as:

$$\mathcal{A}_S = \sum_{k=1}^{N_{\text{view}}} \alpha_k \hat{\mathbf{A}}_k, \quad \boldsymbol{\alpha} \in \Delta_{N_{\text{view}}} \ ,$$

where $\Delta_{N_{\text{view}}}$ denotes a N_{view}-dimensional vector.

After obtaining the representation $\bar{\mathcal{Z}}$ and $\hat{\mathcal{Z}}$ from two streams respectively, we develop a gate mechanism to fuse these spatial features and project the fused feature back to the original patches via a trainable linear projection $\mathbf{W}'_p \in \mathbb{R}^{d \times P}$ as follows:

$$g = \text{Sigmoid}(\bar{\mathcal{Z}} + \hat{\mathcal{Z}}) ,$$
$$\tilde{\mathcal{P}} = \left(g\bar{\mathcal{Z}} + (1-g)\hat{\mathcal{Z}}\right)\mathbf{W}'_p .$$

By concatenating patches on all channels, we obtain $\hat{\mathcal{X}}$ to reconstruct the origin traffic signal $\mathcal{X}$.

Training Strategy and Anomaly Criterion

In addition to the contrastive heterogeneity loss mentioned in Eqn. (1), the loss function of our framework also includes a reconstruction loss and a dual-stream contrastive loss:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \alpha\mathcal{L}_{\text{hetero}} + \beta\mathcal{L}_{\text{dual}} ,$$

where $\alpha, \beta > 0$ are hyper-parameters to control the relative importance of the three objectives.

Our method is an encoder-decoder framework to reconstruct normal data instances accurately, thereby aiming to detect the anomalies as those that fail to accurately reconstruct them under the learned model, that is

$$\mathcal{L}_{\text{recon}} = \frac{1}{N}\sum_{n=1}^{N}\left\|\mathcal{X}^{(n)} - \hat{\mathcal{X}}^{(n)}\right\|_2^2 ,$$

We specifically design a dual-stream contrastive loss to extract the abnormal traffic flows that are inconsistent with their expected service level and functionality. With the dual-stream structure, spatio-temporal dependencies learned from the adaptive data-driven view and stable pre-defined view are captured respectively. The intuition is that the dynamic and static spatio-temporal dependence of an *anomaly region* will lead to large discrepancy in distributions. We formalize a loss function based on Kullback–Leibler divergence (KL divergence) to measure the similarity of the adaptively learned dynamic spatio-temporal dependency $\mathcal{A}_D$ and the static spatio-temporal dependency $\mathcal{A}_S$. Thus, we first design two contrastive loss functions for dynamic global-context stream $\mathcal{L}_D$ and static service-flow stream $\mathcal{L}_S$, respectively, which are defined as follows:

$$\mathcal{L}_D(\mathcal{A}_D, \mathcal{A}_S) = \frac{1}{N}\sum_{n=1}^{N}\textit{sysKL}\big(\mathcal{A}_D(n), \text{Stopgrad}(\mathcal{A}_S(n))\big)$$
$$\mathcal{L}_S(\mathcal{A}_D, \mathcal{A}_S) = \frac{1}{N}\sum_{n=1}^{N}\textit{sysKL}\big(\mathcal{A}_D(n), \text{Stopgrad}(\mathcal{A}_S(n))\big) ,$$

where $\textit{sysKL}(p, q) = (\textit{KL}(p, q) + \textit{KL}(q, p))/2$ and 'Stopgrad' denotes the stop-gradient operation. To train the two branches asynchronously, we formulate the dual-stream contrastive loss as follows:

$$\mathcal{L}_{\text{dual}} = \mathcal{L}_D(\mathcal{A}_D, \mathcal{A}_S) - \mathcal{L}_S(\mathcal{A}_D, \mathcal{A}_S) .$$

Table 1: Statistics of Datasets.

Dataset	# Nodes	# Edges	Time interval	# Anomalies
PeMS03	358	546	5 min	36
PeMS08	170	277	5 min	17
NYC-Taxi-Yellow	263	1097	10 min	26
NYC-Taxi-Green	263	1097	10 min	26

After the loss convergence, the anomaly score for each node/region will depend on both the reconstruction error and the contrastive losses:

$$\text{Score}(v_n) = \|\mathcal{X}^{(n)} - \hat{\mathcal{X}}^{(n)}\|_2^2 + \beta \cdot \textit{sysKL}(\mathcal{A}_D(n), \mathcal{A}_S(n))$$

A hyper-parameter threshold ϵ is used with the anomaly scores to detect long-term traffic anomalies.

Experiments

Benchmark Datasets

To thoroughly evaluate our proposed model, we conduct extensive experiments on 4 real-world traffic flow and speed benchmark datasets:

(1) PEMS03 and PEMS08 (Fang et al. 2023). These datasets are collected from the California Transportation Performance Measurement System (PeMS). We also contain distance and connectivity information between different transportation districts. The traffic flow data has a fine-grained time resolution of 5 minutes between consecutive time steps.

(2) NYC-Yellow-Taxi and NYC-Green-Taxi dataset (Yao et al. 2019), which contains inflow and outflow data of yellow taxis and green taxis in the New York City from 01/01/2023 to 02/31/2023. We divide the regions and construct them as graph structures. The multi-view adjacency matrices are built to include distance, connectivity, and POI similarity matrix according to taxi zones. We set the time interval as 10 minutes.

Inspired by previous work (Ding et al. 2019; Ding, Li, and Liu 2019), we proposed two long-term traffic anomaly injection methods: long-term temporal anomaly injection and long-term spatial anomaly injection. For the temporal anomaly injection, we randomly choose regions and simulate periodic traffic jams caused by the lack of regional service facilities. Formally, given a threshold $\mu \in (0, 1)$ and the traffic signal $\mathbf{X}^{(n)}$ of the node v_n, the flow after injection is denoted as follows,

$$\mathbf{X}^{(n)} \leftarrow \min(\mathbf{X}^{(n)}, \mu \max(\mathbf{X}^{(n)})) .$$

In addition to the injection of temporal anomalies, to inject spatial anomalies, we adopt a signal perturbation schema introduced by (Song et al. 2007), which is a popular anomaly injection method in graph node anomaly detection:

$$\mathbf{X}^{(n)} \leftarrow \mathbf{X}^{(i)},\ \text{where } i = \underset{j \in \text{Random}(N,k)}{\arg\max} \|\mathbf{X}^{(n)} - \mathbf{X}^{(k)}\|_2 ,$$

and Random(N, k) denotes randomly sampling k elements from the set $\{1, 2, \cdots, N\}$. In our experiments, we set the value of k to 10% of the number of nodes. The detailed statistics of datasets are summarized in Table 1.

Table 2: Overall results on real-world traffic datasets (%). The best ones are in **Bold**.

Dataset	**PeMS03**			**PeMS08**			**NYC-Yellow-Taxi**			**NYC-Green-Taxi**		
Metric	**Recall@5%**	**Recall@10%**	**ROC-AUC**	**Recall@5%**	**Recall@10%**	**ROC-AUC**	**Recall@5%**	**Recall@10%**	**ROC-AUC**	**R5%**	**Recall@10%**	**ROC-AUC**
LOF	10.00	17.59	41.78	6.25	18.84	33.40	12.32	20.85	39.20	11.37	18.96	32.12
OCSVM	20.59	26.47	48.51	18.75	31.25	47.10	50.00	57.69	75.55	34.62	38.46	71.04
DOMINANT	23.53	38.24	58.24	25.00	32.50	77.79	3.85	5.77	91.15	3.85	21.54	74.88
AnomalyDAE	23.53	38.24	67.95	2.50	27.50	59.57	3.85	5.38	86.54	3.85	19.23	64.11
OCGNN	21.76	38.24	52.41	6.25	31.25	44.47	3.85	7.30	83.76	3.85	19.23	69.53
LSTM-VAE	23.53	38.24	54.91	5.00	31.25	75.93	3.85	7.69	84.13	3.85	19.23	75.65
AnomalyTrans	18.82	36.47	64.28	23.75	42.50	81.44	35.38	69.23	83.10	26.92	54.62	75.77
STDecomp	23.53	38.24	52.98	6.25	31.25	78.81	30.76	42.31	79.88	3.85	19.23	64.48
STGAN	17.65	38.24	61.34	6.25	31.25	49.80	3.85	7.69	52.67	3.85	19.23	53.02
ours DSCL	**43.52**	**84.71**	**86.88**	**28.75**	**65.00**	**83.66**	**46.92**	**96.15**	**96.22**	**40.77**	**74.62**	**86.40**

Baseline and Evaluation Criteria

In this section, we introduce the detailed experimental settings, including the baseline methods and evaluation metrics. We compare the proposed DSCL framework with nine baselines for comprehensive evaluations, which can be divided into four groups:

- Individual Methods: LOF (Breunig et al. 2000) and OCSVM (Schölkopf et al. 2001);
- Graph anomaly detection methods: DOMINANT (Ding et al. 2019), AnomalyDAE (Fan, Zhang, and Li 2020) and OCGNN (Wang et al. 2021);
- Time-series anomaly detection methods: LSTM-VAE (Lin et al. 2020) and AnoamlyTrans (Xu et al. 2022);
- Spatio-temporal anomaly detection methods: STDecomp (Zhang et al. 2019) and STGAN (Deng et al. 2022).

For time-series and short-term spatio-temporal anomaly detection methods, we take the sum of the anomaly scores on all the timestamps of a node as its anomaly score. We adopt Area under the ROC Curve (ROC-AUC) and recall at top-k positions (Recall@k) as evaluation metrics to measure the performance of anomaly detection algorithms.

Implementation Details

In the experiments, we use the min-max normalization method to scale the traffic data into the range $[0, 1]$. The time length of training and testing are both set to 14 days for each dataset. For PeMS datasets, distance, and connectivity information is provided, i.e. $N_{\text{view}} = 2$; for the NYC-Taxi dataset, we obtain its distance, connectivity, and POI information, thus three views are provided to DSCL.

For the DSCL model, the dimension of the hidden state is 128, the number of temporal attention heads is 16, and the number of spatial attention heads is 8. The number of layers in scalable multi-view GCN is set to 3. The patch sizes of different datasets are all set to 12. All the experiments are implemented in PyTorch with one NVIDIA 3060 12GB GPU. Adam (Kingma and Ba 2014) with default parameter applied for optimization. We set the initial learning rate to 10^{-3} with 500 epochs for all datasets. We conduct experiments with 10 independent runs.

Anomaly Detection Results

We first evaluate our DSCL with 10 competitive baselines on 4 real-world traffic datasets, as shown in Table 2. From the evaluation results, we make the following observations: (1) The proposed DSCL method achieves SOTA results under the widely used ROC-AUC (Ding et al. 2019; Fan, Zhang, and Li 2020) and Recall@k (Deng et al. 2022) in most benchmark datasets. It verifies the effectiveness of the dual-stream spatio-temporal contrastive learning framework. (2) The individual methods cannot achieve satisfying results in most datasets as they merely consider the nodal traffic signals. The attributed graph anomaly detection methods help to detect nodes where traffic signals have incompatibility with their expected service levels. However, they do not resolve the long-range temporal dependencies in traffic signals. The time-series anomaly detection method lacks the consideration of spatial dependencies between nodes. (3) Short-term traffic anomaly detection methods, such as STDecomp and STGAN, focus on identifying traffic flow deviations within specific time slots. These approaches face challenges in detecting recurrent anomalies, as their reconstruction models may classify such occurrences as normal patterns. Furthermore, relying solely on the cumulative occurrence of short-term anomalies demonstrates that these methods have limitations in effectively discerning potential long-term spatial-temporal anomalies. (4) Note that in the long-term traffic anomaly detection problem, some baseline methods show high ROC-AUCs with low recalls, which means they mistakenly treat the normal areas as anomalies, resulting in a waste of urban construction resources.

Ablation Study

To further investigate the effectiveness of the key components in DSCL, we conduct ablation studies on all datasets. By disabling different components, we obtained the performance results of the variants as below.

Effect of stream We present ROC-AUC values on the four datasets with two variants: (1) *w/o Static*: this variant removes the static service-flow stream. (2) *w/o Dynamic*: this variant removes the dynamic global-context stream. Table 3 shows the comparison of these variants.

The results show that both the streams are important to detect long-term spatial-temporal traffic anomalies.The results on PeMS datasets indicate that the proposed dual-stream structure dramatically increases the detection performance for long-term traffic anomalies. The results in the NYC Taxi datasets also demonstrate that only using the regular reconstruction methods is not effective enough for our task. The designed multi-view scalable GCN in the static service-flow stream guarantee the extraction of semantic-aware correlations with the expected serviceability of regions.

Table 3: Ablation studies on stream setting (%).

Stream setting	PeMS03	PeMS08	NYC-Yellow-Taxi	NYC-Green-Taxi
w/o Static	54.84	39.80	86.09	83.36
w/o Dynamic	49.66	37.86	89.60	83.57
DCSL	86.88	83.66	96.15	86.40

Effect of anomaly detector In this part, we study the effect of our designed long-term anomaly detector in DCSL. We discuss the performance (ROC-AUC values) of 4 variants of the proposed model: (1) *w/o* $\mathcal{L}_{\text{hetero}}$&$\mathcal{L}_{\text{dual}}$: this variant trains the model with reconstruction loss only. (2) *w/o* $\mathcal{L}_{\text{hetero}}$&$\mathcal{L}_{\text{recon}}$: this variant trains the model with dual-stream contrastive loss only. (3) *w/o* $\mathcal{L}_{\text{dual}}$: this variant trains the model with reconstruction loss and contrastive heterogeneous loss. (4) *w/o* $\mathcal{L}_{\text{hetero}}$: this variant trains the model with reconstruction loss and dual-stream contrastive loss. (5) *w/o* $\mathcal{L}_{\text{recon}}$: this variant trains the model with dual-stream contrastive loss and contrastive heterogeneous loss.

As shown in Table 4, the experimental results demonstrate that both reconstruction loss and dual-stream contrastive loss play vital roles in detecting long-term traffic anomalies. The importance of each depends on the specific datasets. It is worth to be noted that contrastive heterogeneity loss makes the model perform more comprehensive detection. It demonstrated that proposed soft contrastive clustering method can effectively distinct different types of traffic patterns of regions caused by functionalities and guarantee spatial self-attention block to learn comprehensive dynamic global-context dependencies.

Table 4: Ablation studies on loss setting (%).

Loss Setting	PeMS03	PeMS08	NYC-Yellow-Taxi	NYC-Green-Taxi
w/o $\mathcal{L}_{\text{hetero}}$&$\mathcal{L}_{\text{dual}}$	51.84	38.05	92.73	83.77
w/o $\mathcal{L}_{\text{hetero}}$&$\mathcal{L}_{\text{recon}}$	66.70	77.43	43.43	48.40
w/o $\mathcal{L}_{\text{dual}}$	57.09	57.84	95.14	85.11
w/o $\mathcal{L}_{\text{hetero}}$	77.12	78.37	94.66	86.10
w/o $\mathcal{L}_{\text{recon}}$	79.96	80.20	51.43	53.79
DCSL	86.88	83.66	96.15	86.40

Case Study

In this section, we implement our method on the real NYC-Yellow-Taxi dataset without anomaly injection. Our goal is to detect long-term anomalies in current urban traffic. The detection result is presented in Figure 3.

The result reveals that the long-term anomalies mainly happen in the center of Manhattan. We observed that the recurrent traffic congestion usually happened in the evening around Times Square and Lincoln Center for the Performing Arts. As shown in Figure 4, we observed that their traffic demands are dramatically higher than those places with similar functionality and their own regular time, leading to recurrent traffic anomalies. We analyze that during our collected data(01/01/2023 to 02/31/2023), these two places held many important celebrations and events. In summary, our method can

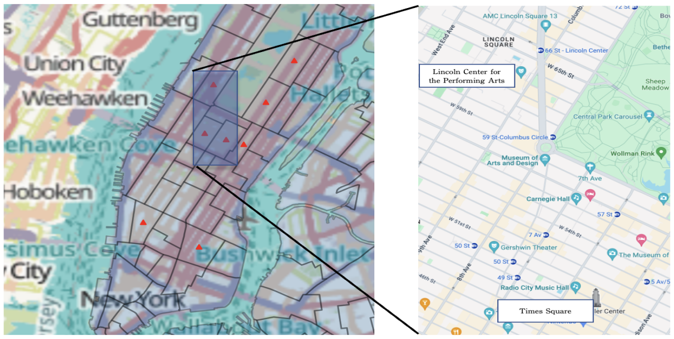


Figure 3: Anomaly detection result on NYC-Yellow-Taxi data. The regions with the highest anomaly scores are marked with red marks.

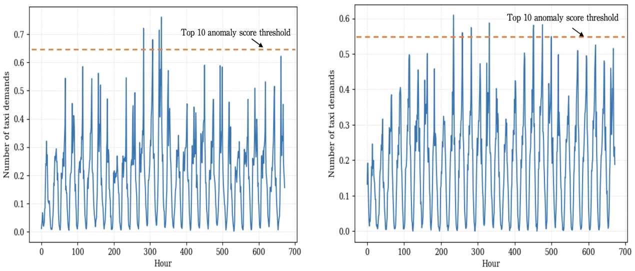


Figure 4: An example of the detected anomalies in NYC.

effectively detect long-term traffic anomalies, which could help urban transportation planning. The existing short-term traffic anomaly detection methods are insufficient for this task since they are unable to effectively learn long and stable spatio-temporal patterns and extract recurrent patterns.

Conclusion

In this paper, we propose a novel Dual-stream Spatio-temporal Contrastive Learning framework, called DSCL, for long-term traffic anomaly detection. We propose two stream learning networks to extract multiple complex correlations of traffic signals from urban global-context information among regions and stable expected traffic capabilities. The dynamic stream consists of novel spatio-temporal attention blocks and a contrastive soft clustering algorithm, which helps to extract region heterogeneity adaptively. In the static stream, we propose a novel scale multi-view GCN with a temporal attention block to learn the traffic pattern similarity and heterogeneity from the relationships with corresponding services. DSCL also integrates the traffic graph reconstruction loss and two contrastive losses to achieve efficient traffic anomaly detection. Extensive experiments conducted on the current real-world case of traffic systems demonstrate the superiority of the DSCL over the other baseline methods. In the future, we will further explore this problem to detect both short-term and long-term traffic anomalies based on a single unified model.

References

Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; and Sander, J. 2000. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 93–104.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*.

Cho, H.; Seol, J.; and goo Lee, S. 2023. Masked Contrastive Learning for Anomaly Detection. arXiv:2105.08793.

Chow, A. H. F.; Santacreu, A.; Tsapakis, I.; Tanasaranond, G.; and Cheng, T. 2014. Empirical assessment of urban traffic congestion. *Journal of Advanced Transportation*, 48: 1000–1016.

Deng, L.; Lian, D.; Huang, Z.; and Chen, E. 2022. Graph convolutional adversarial networks for spatiotemporal anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6): 2416–2428.

Ding, K.; Li, J.; Bhanushali, R.; and Liu, H. 2019. Deep anomaly detection on attributed networks. In *Proceedings of the International Conference on Data Mining*, 594–602.

Ding, K.; Li, J.; and Liu, H. 2019. Interactive anomaly detection on attributed networks. In *Proceedings of the ACM international conference on web search and data mining*.

Fan, H.; Zhang, F.; and Li, Z. 2020. Anomalydae: Dual autoencoder for anomaly detection on attributed networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5685–5689. IEEE.

Fang, Y.; Qin, Y.; Luo, H.; Zhao, F.; Xu, B.; Zeng, L.; and Wang, C. 2023. When spatio-temporal meet wavelets: Disentangled traffic forecasting via efficient spectral graph attention networks. In *IEEE 39th International Conference on Data Engineering*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

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

Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.-X.; and Yan, X. 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems*, 32.

Lin, C.; Zhu, Q.; Guo, S.; Jin, Z.; Lin, Y.-R.; and Cao, N. 2018. Anomaly detection in spatiotemporal data via regularized non-negative tensor analysis. *Data Mining and Knowledge Discovery*, 32: 1056–1073.

Lin, S.; Clark, R.; Birke, R.; Schönborn, S.; Trigoni, N.; and Roberts, S. 2020. Anomaly detection for time series using vae-lstm hybrid model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Liu, Y.; Li, Z.; Pan, S.; Gong, C.; Zhou, C.; and Karypis, G. 2021. Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE transactions on neural networks and learning systems*, 33(6): 2378–2392.

Mihaita, A.-S.; Li, H.; and Rizoiu, M.-A. 2020. Traffic congestion anomaly detection and prediction using deep learning. *arXiv preprint arXiv:2006.13215*.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. arXiv:2211.14730.

Schölkopf, B.; Platt, J. C.; Shawe-Taylor, J.; Smola, A. J.; and Williamson, R. C. 2001. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7): 1443–1471.

Song, X.; Wu, M.; Jermaine, C.; and Ranka, S. 2007. Conditional anomaly detection. *IEEE Transactions on knowledge and Data Engineering*, 19(5): 631–645.

Wang, J.; Wu, J.; Wang, Z.; Gao, F.; and Xiong, Z. 2019. Understanding urban dynamics via context-aware tensor factorization with neighboring regularization. *IEEE Transactions on Knowledge and Data Engineering*, 32(11): 2269–2283.

Wang, X.; Jin, B.; Du, Y.; Cui, P.; Tan, Y.; and Yang, Y. 2021. One-class graph neural networks for anomaly detection in attributed networks. *Neural computing and applications*, 33: 12073–12085.

Witayangkurn, A.; Horanont, T.; Sekimoto, Y.; and Shibasaki, R. 2013. Anomalous event detection on large-scale gps data from mobile phones using hidden markov model and cloud platform. In *Proceedings of the ACM conference on Pervasive and ubiquitous computing adjunct publication*.

Xu, J.; Wu, H.; Wang, J.; and Long, M. 2022. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy. In *International Conference on Learning Representations*.

Xu, M.; Dai, W.; Liu, C.; Gao, X.; Lin, W.; Qi, G.-J.; and Xiong, H. 2020. Spatial-temporal transformer networks for traffic flow forecasting. *arXiv preprint arXiv:2001.02908*.

Yang, Y.; Zhang, C.; Zhou, T.; Wen, Q.; and Sun, L. 2023. DCdetector: Dual Attention Contrastive Representation Learning for Time Series Anomaly Detection. *arXiv preprint arXiv:2306.10347*.

Yao, H.; Tang, X.; Wei, H.; Zheng, G.; and Li, Z. 2019. Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 5668–5675.

Zhang, H.; Zheng, Y.; and Yu, Y. 2018. Detecting urban anomalies using multiple spatio-temporal data sources. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 2(1): 1–18.

Zhang, M.; Li, T.; Shi, H.; Li, Y.; and Hui, P. 2019. A decomposition approach for urban anomaly detection across spatiotemporal data. In *IJCAI International Joint Conference on Artificial Intelligence*.

Zhang, M.; Li, T.; Yu, Y.; Li, Y.; Hui, P.; and Zheng, Y. 2020. Urban anomaly analytics: Description, detection, and prediction. *IEEE Transactions on Big Data*, 8(3): 809–826.

Reproducibility Checklist

Unless specified otherwise, please answer "yes" to each question if the relevant information is described either in the paper itself or in a technical appendix with an explicit reference from the main paper. If you wish to explain an answer further, please do so in a section titled "Reproducibility Checklist" at the end of the technical appendix.

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced. Yes.
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results. Yes.
- Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper. Yes.

Does this paper make theoretical contributions? Yes.
If yes, please complete the list below.

- All assumptions and restrictions are stated clearly and formally. Yes.
- All novel claims are stated formally (e.g., in theorem statements). Yes.
- Proofs of all novel claims are included. Tes.
- Proof sketches or intuitions are given for complex and/or novel results. Yes.
- Appropriate citations to theoretical tools used are given. Yes.
- All theoretical claims are demonstrated empirically to hold. Yes.
- All experimental code used to eliminate or disprove claims is included. Yes.

Does this paper rely on one or more datasets? Yes.
If yes, please complete the list below.

- A motivation is given for why the experiments are conducted on the selected datasets. Yes.
- All novel datasets introduced in this paper are included in a data appendix. Yes.
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. Yes.
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. Yes.
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. Yes.
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. Yes.

Does this paper include computational experiments? Yes.
If yes, please complete the list below.

- Any code required for pre-processing data is included in the appendix. Yes.
- All source code required for conducting and analyzing the experiments is included in a code appendix. Yes.
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. Yes.
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from. Yes.
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. Yes.
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. Yes.
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. Yes.
- This paper states the number of algorithm runs used to compute each reported result. Yes.
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. Yes.
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). Yes.
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. Yes.
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. Yes.