Temporal Multi-view Graph Convolutional Networks for Citywide Traffic Volume Inference

Shaojie Dai, Jinshuai Wang, Chao Huang, Yanwei Yu* and Junyu Dong

Ocean University of China

Email: daishaojie@stu.ouc.edu.cn

Outline

- **Background**
- Problem
- Framework
- Experiment
- Conclusion

Background

Citywide traffic volume inference is key to an intelligent city.

Intelligent transportation system

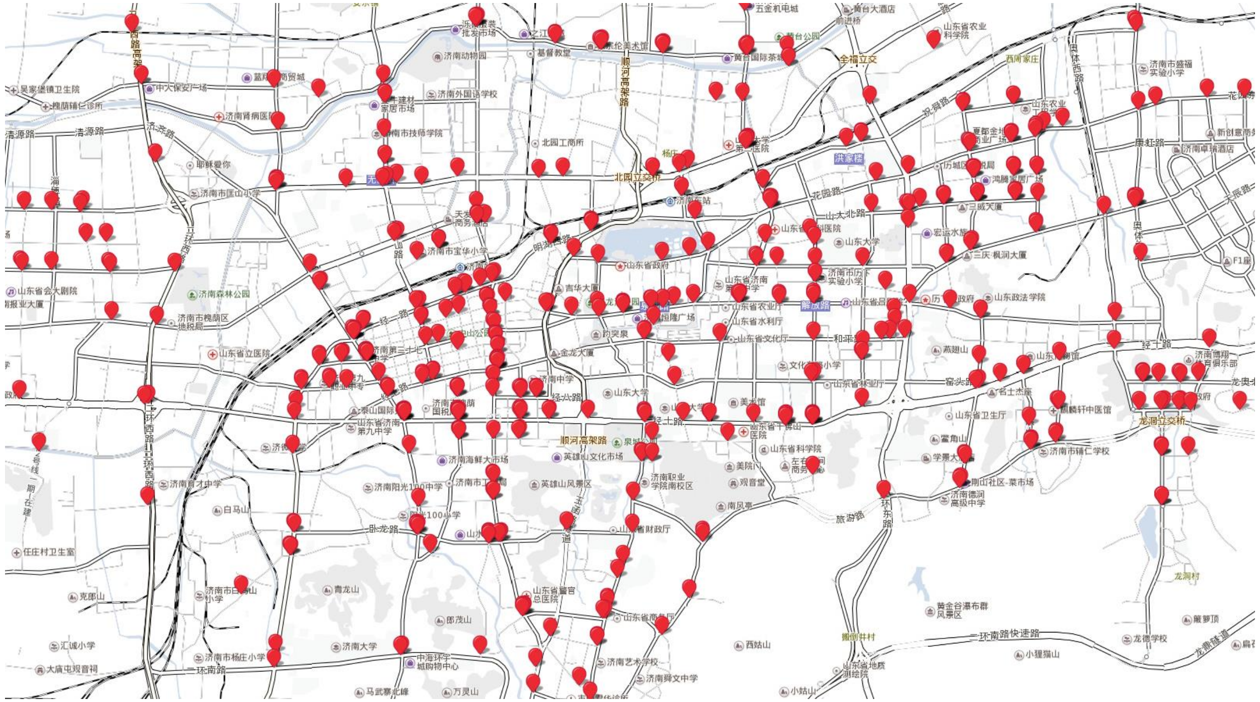
Alleviating traffic congestion

Government's policy-making

Background

Citywide traffic volume inference is a **challenging** task because:

- **Coverage is limited**: accurate traffic volumes on the roads can only be measured at certain locations where **sensors** are installed.



Only 2% of road segments in Jinan city deploy the surveillance cameras for traffic monitoring [1].

[1] X. Yi, Z. Duan, T. Li, T. Li, J. Zhang, and Y. Zheng, "Citytraffic: Modeling citywide traffic via neural memorization and generalization approach," inCIKM, 2019, pp. 2665–2671.

Background

Citywide traffic volume inference is a **challenging** task because:

- **Lack of Historical Observations:** it is worth noting that **different** from the problem of traffic volume **forecast** based on the historical data, there is no any historical data available for the unmonitored roads.

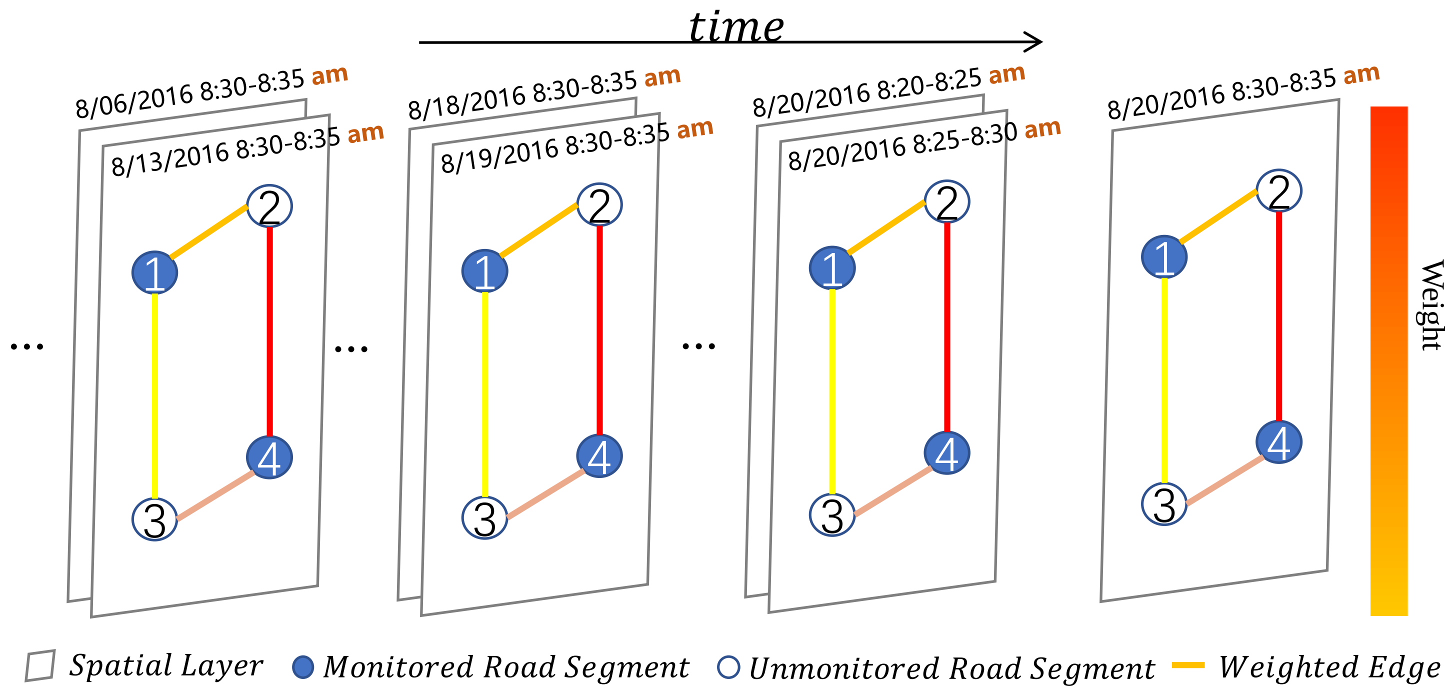
Volume(vehicle/5 minute/road)

$$
\begin{array}{c}
\\
t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_m
\end{array}
\begin{array}{c}
\begin{array}{cccccc} r_1 & r_2 & r_3 & \cdots & r_{n-1} & r_n \end{array} \\
\left[\begin{array}{cccccc}
? & 13 & ? & \cdots & ? & 24 \\
? & 18 & ? & \cdots & ? & 29 \\
? & 20 & ? & \cdots & ? & 37 \\
 & & & & & \\
? & 15 & ? & \cdots & ? & 18
\end{array}\right]
\end{array}
$$

Background

Citywide traffic volume inference is a **challenging** task because:

- **Complex Spatial-Temporal Dependencies:** Different granularity-specific variation regularities of traffic data may present various temporal patterns (e.g., hourly, daily, weekly) which are complementary and inter-dependent with each other [2].

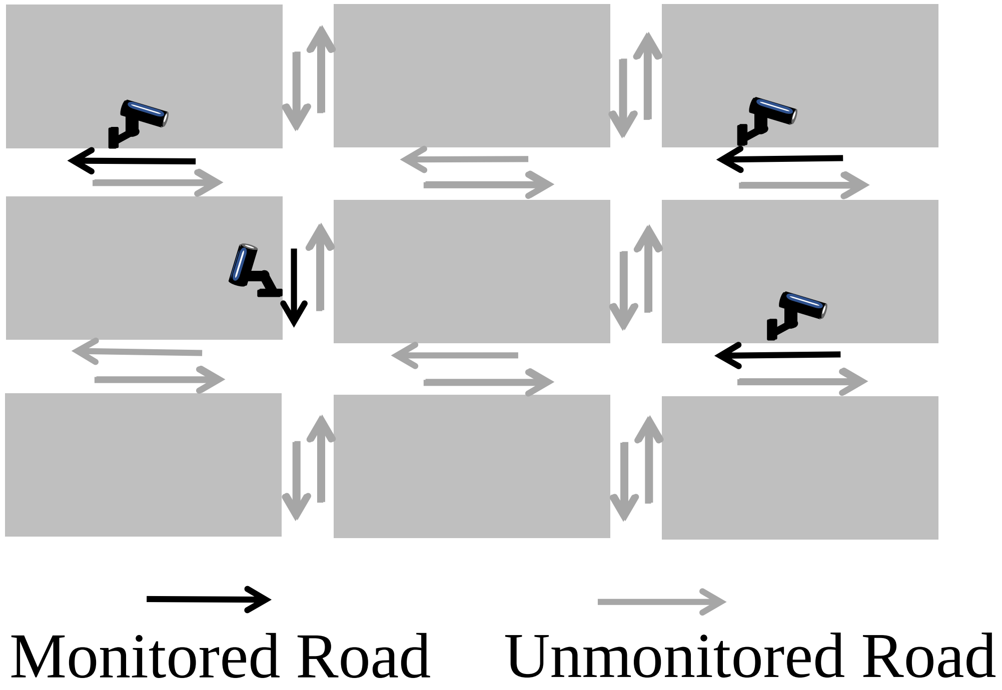


[2] Y. Yu, X. Tang, H. Yao, X. Yi, and Z. Li, "Citywide traffic volume inference with surveillance camera records," IEEE Transactions on Big Data, 2019.

Outline

- Background
- **Problem**
- Framework
- Experiment
- Conclusion

Problem



$$
\begin{array}{c}
 \\ t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_m
\end{array}
\begin{array}{c}
\begin{array}{cccccc} r_1 & r_2 & r_3 & \cdots & r_{n-1} & r_n \end{array} \\
\begin{bmatrix}
? & 13 & ? & \cdots & ? & 24 \\
? & 18 & ? & \cdots & ? & 29 \\
? & 20 & ? & \cdots & ? & 37 \\
 & & & & & \\
? & 15 & ? & \cdots & ? & 18
\end{bmatrix}
\end{array}
$$

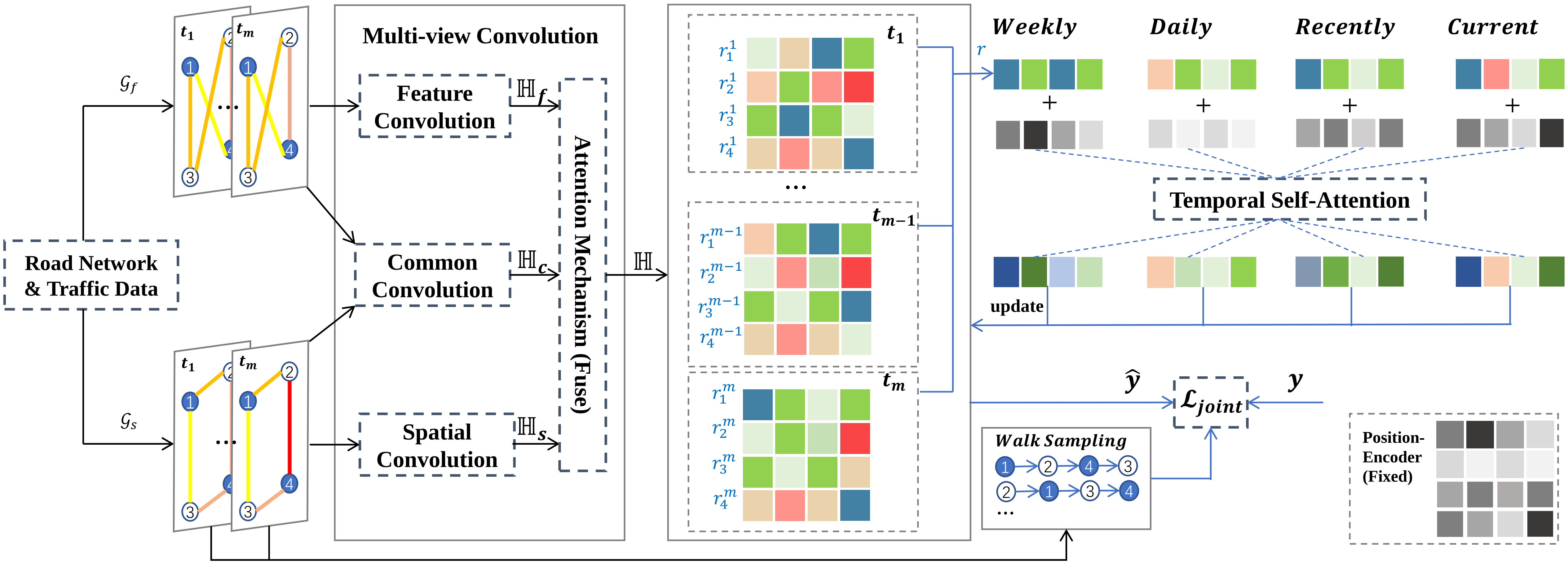
Volume(vehicle/5 minute/road)

Given a road network, observed traffic volume at the monitored road segments, our goal is to infer citywide traffic volume of any unmonitored road segment, $r_i \in \mathcal{U}$, at any time interval. ($\boldsymbol{n}$ road segment, $\boldsymbol{m}$ time slice)

Outline

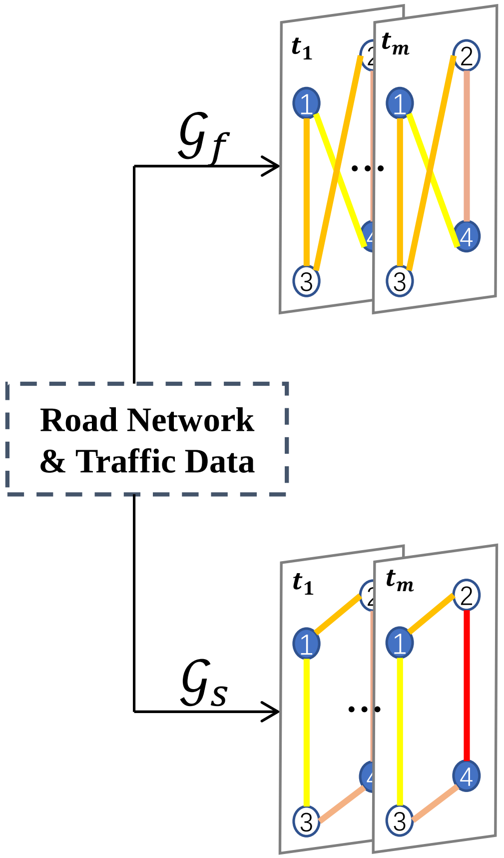
- Background
- Problem
- **Framework**
- Experiment
- Conclusion

Framework



Road Network & Traffic Data
$\mathcal{G}_f$
$\mathcal{G}_s$
t_1
t_m
Multi-view Convolution
Feature Convolution
Common Convolution
Spatial Convolution
$\mathbb{H}_f$
$\mathbb{H}_c$
$\mathbb{H}_s$
Attention Mechanism (Fuse)
$\mathbb{H}$
t_{m-1}
r_1^1 r_2^1 r_3^1 r_4^1
r_1^{m-1} r_2^{m-1} r_3^{m-1} r_4^{m-1}
r_1^m r_2^m r_3^m r_4^m
r
Weekly
Daily
Recently
Current
Temporal Self-Attention
update
$\hat{y}$
$\mathcal{L}_{joint}$
y
Walk Sampling
Position-Encoder (Fixed)

Affinity Graph Construction



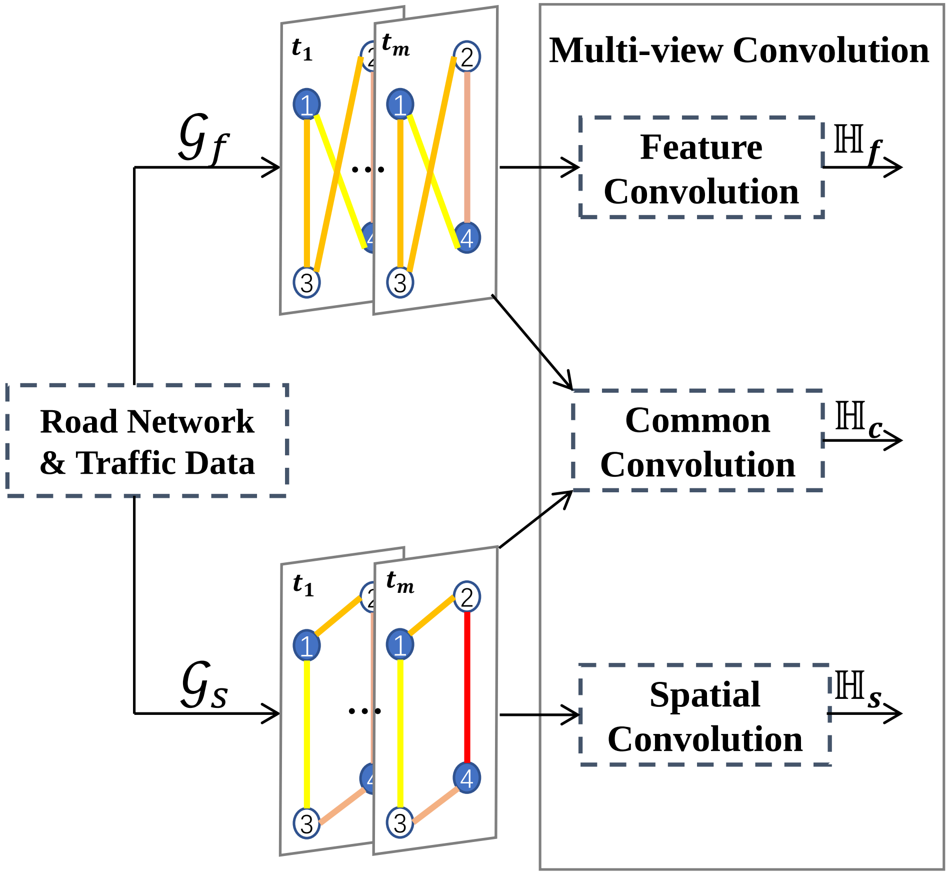
$$s_{ij} = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{|\mathbf{x}_i||\mathbf{x}_j|}$$

$\mathbf{x}_i$ denotes the feature of road segment i.

$$w_{ij} = \sigma(liner(\frac{max(lane_i, lane_j)}{min(lane_i, lane_j)}))$$

$lane_i$ denotes the #lanes of road segment i.

Multi-view Graph Convolution Network



$$\mathbf{H}_f^{(l+1)} = Relu(\widetilde{\mathbf{D}}_f^{-\frac{1}{2}}\widetilde{\mathbf{A}}_f\widetilde{\mathbf{D}}_f^{-\frac{1}{2}}\boldsymbol{H}_f^{(l)}\mathbf{W}_f^{(l)})$$

$$\mathbf{H}_{cf}^{(l+1)} = Relu(\widetilde{\mathbf{D}}_f^{-\frac{1}{2}}\widetilde{\mathbf{A}}_f\widetilde{\mathbf{D}}_f^{-\frac{1}{2}}\boldsymbol{H}_{cf}^{(l)}\mathbf{W}_c^{(l)})$$

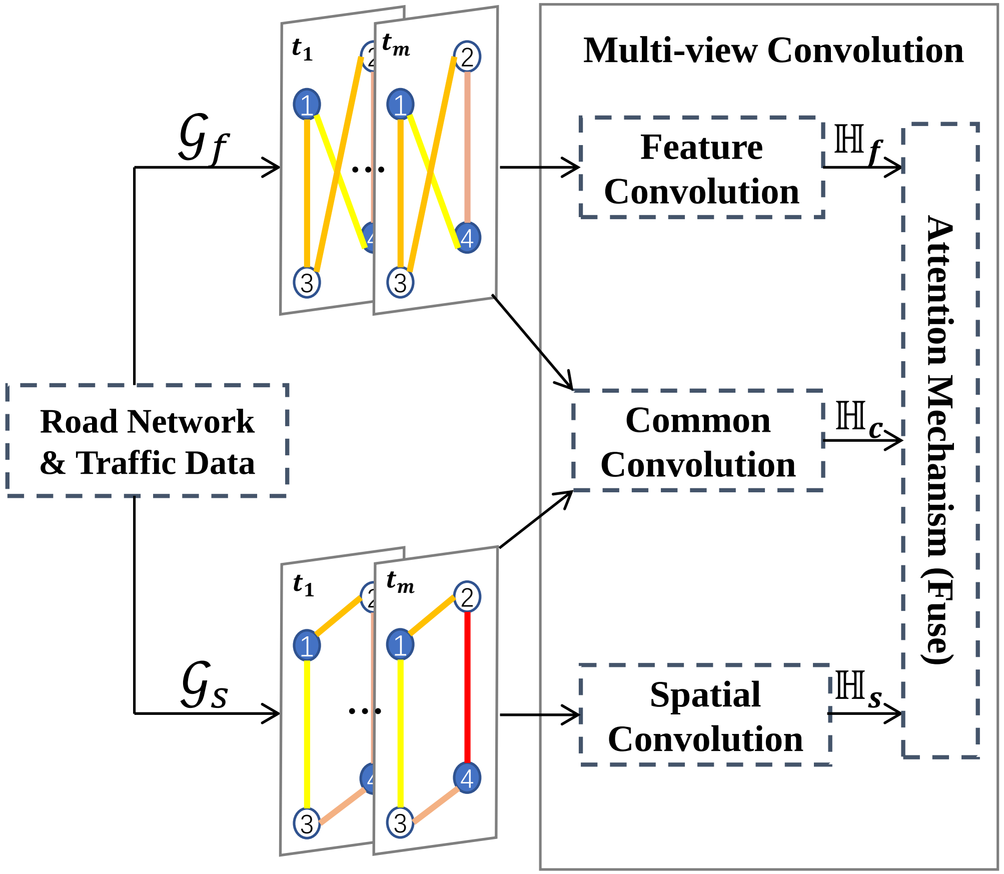
$$\mathbf{H}_{cs}^{(l+1)} = Relu(\widetilde{\mathbf{D}}_s^{-\frac{1}{2}}\widetilde{\mathbf{A}}_s\widetilde{\mathbf{D}}_s^{-\frac{1}{2}}\boldsymbol{H}_{cs}^{(l)}\mathbf{W}_c^{(l)})$$

$$\mathbf{H}_c^{(l)} = \frac{\mathbf{H}_{cf}^{(l)} + \mathbf{H}_{cs}^{(l)}}{2}$$

$$\mathbf{H}_s^{(l+1)} = Relu(\widetilde{\mathbf{D}}_s^{-\frac{1}{2}}\widetilde{\mathbf{A}}_s\widetilde{\mathbf{D}}_s^{-\frac{1}{2}}\boldsymbol{H}_s^{(l)}\mathbf{W}_s^{(l)})$$

[3] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in ICLR, 2016.

Multi-view Graph Convolution Network



We finally utilize the attention mechanism $\mathbf{H} = att(\mathbf{H}_s, \mathbf{H}_f, \mathbf{H}_c)$[3] to combine their embedding in a reasonable way as follows:

$$\omega_s^i = \mathbf{q}^\top Tanh\left(\mathbf{W} \cdot \left(\mathbf{h}_s^i\right)^\top + \mathbf{b}\right)$$

$$a_s^i = softmax\left(\omega_s^i\right) = \frac{\exp(\omega_s^i)}{\exp\left(\omega_s^i\right) + \exp\left(\omega_s^i\right) + \exp\left(\omega_s^i\right)}$$

$$\mathbf{a}_S = diag(a_s), \mathbf{a}_F = diag\left(a_f\right), \mathbf{a}_C = diag(a_c)$$

$$\mathbf{H} = \mathbf{a}_S \cdot \mathbf{H}_s + \mathbf{a}_F \cdot \mathbf{H}_f + \mathbf{a}_C \cdot \mathbf{H}_c \quad \mathbf{H} \in R^{n*d}$$

$$\mathbb{H} \in R^{m*n*d}$$

[4] X. Wang, M. Zhu, D. Bo, P. Cui, C. Shi, and J. Pei, “Am-gcn: Adaptive multi-channel graph convolutional networks,” in KDD, 2020, pp. 1243–1253.

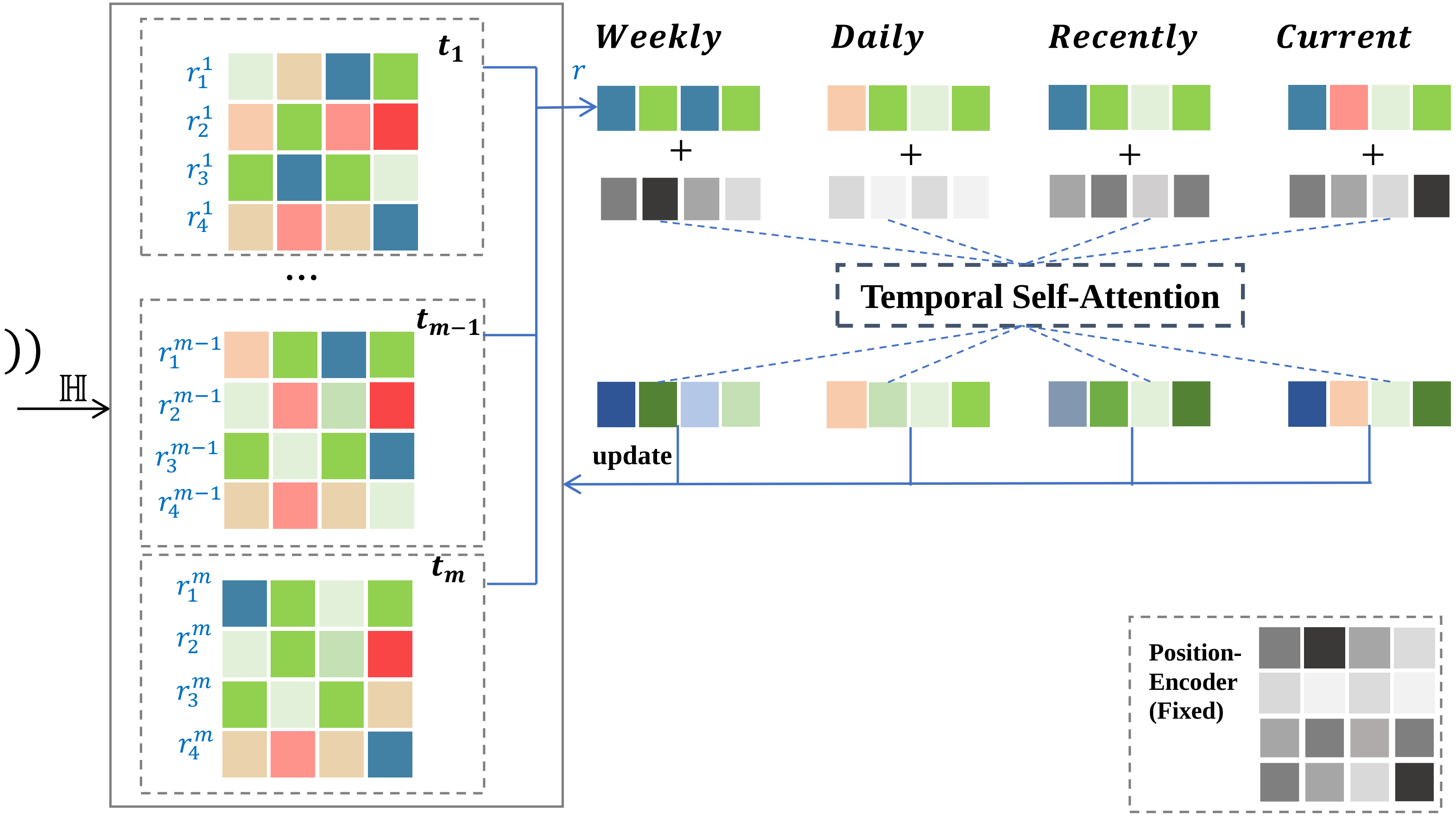
Temporal Self-Attention

$$\mathbf{S}_i = (\mathbf{H}_i + \mathbf{P})\,\mathbf{W}^Q((\mathbf{H}_i + \mathbf{P})\mathbf{W}^K)^\top$$

$$\mathbf{Z}_i = softmax(\frac{\mathbf{S}_i}{\sqrt{d}})(\mathbf{H}_i + \mathbf{P})\,\mathbf{W}^V$$

$$\mathbf{Z}_i = FC(concat(\mathbf{Z}_i^{(1)}, \mathbf{Z}_i^{(2)}, \dots, \mathbf{Z}_i^{(\#head)}))$$

$$\mathbf{P}_{ij} = \begin{cases} \sin\left(\frac{i}{10000^{j/d}}\right) & if\ i\%2 = 0, \\ \cos\left(\frac{i}{10000^{j-1/d}}\right) & else, \end{cases}$$

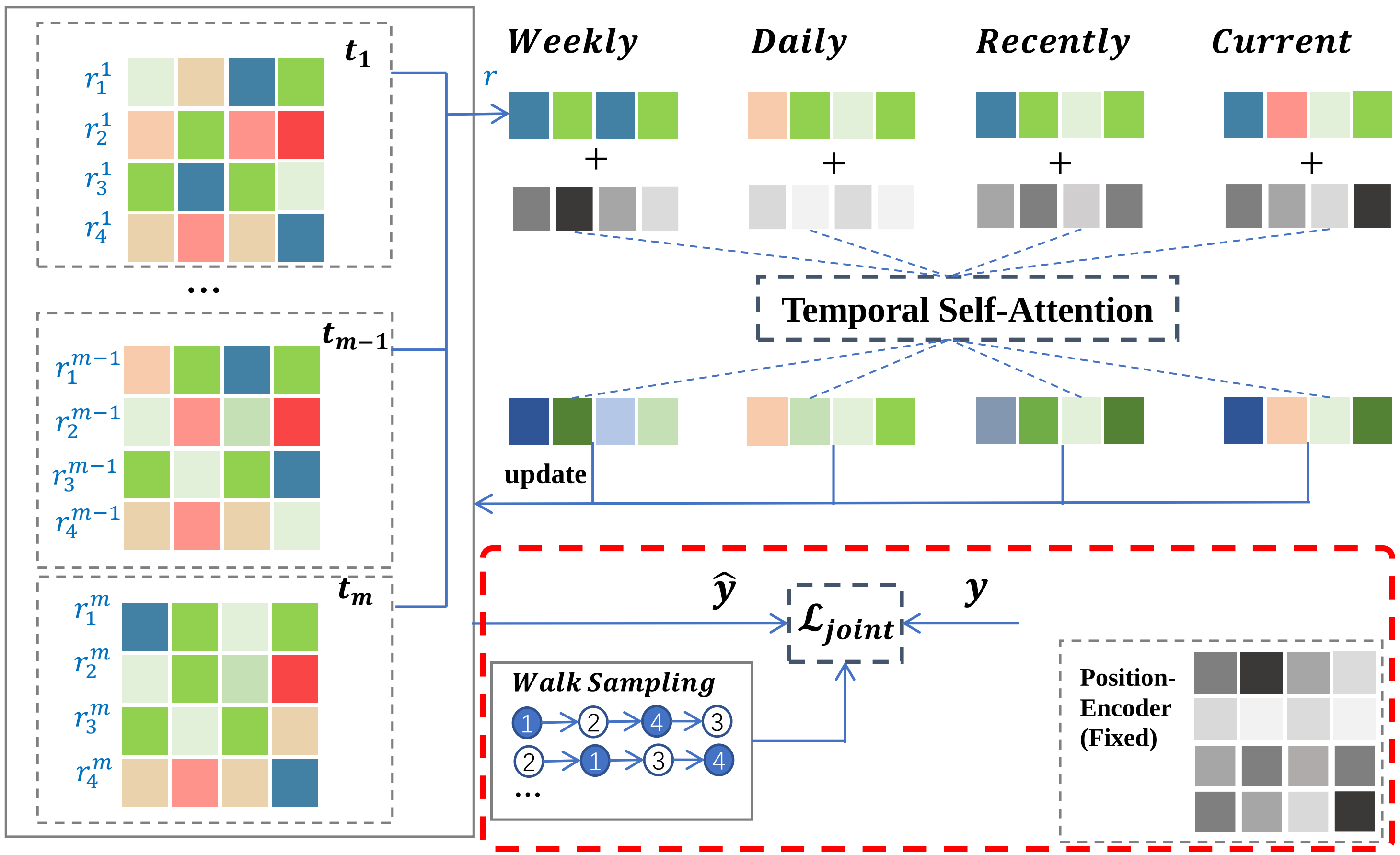


Where $\mathbf{H}_i$ denotes the concatenated hidden representation of road segment r_i at all related time intervals.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,L. Kaiser, and I. Polosukhin, “Attention is all you need,”arXiv preprintarXiv:1706.03762, 2017.

Joint Learning Optimization

How can we learn and make inference?



Joint Learning Optimization

Unsupervised objective function

$$\mathcal{L}_{walk} = \sum_{t \in T} \sum_{v_i \in \mathcal{V}} \Big(\sum_{v_j \in \mathcal{N}_{walk}^{t}(v_i)} - \log\big(\sigma(s_{ij}^{t})\big) - \sum_{v_k \in Neg^{t}(v_i)} \log(1 - \sigma(s_{ij}^{t}))\Big)$$

Semi-supervised objective function

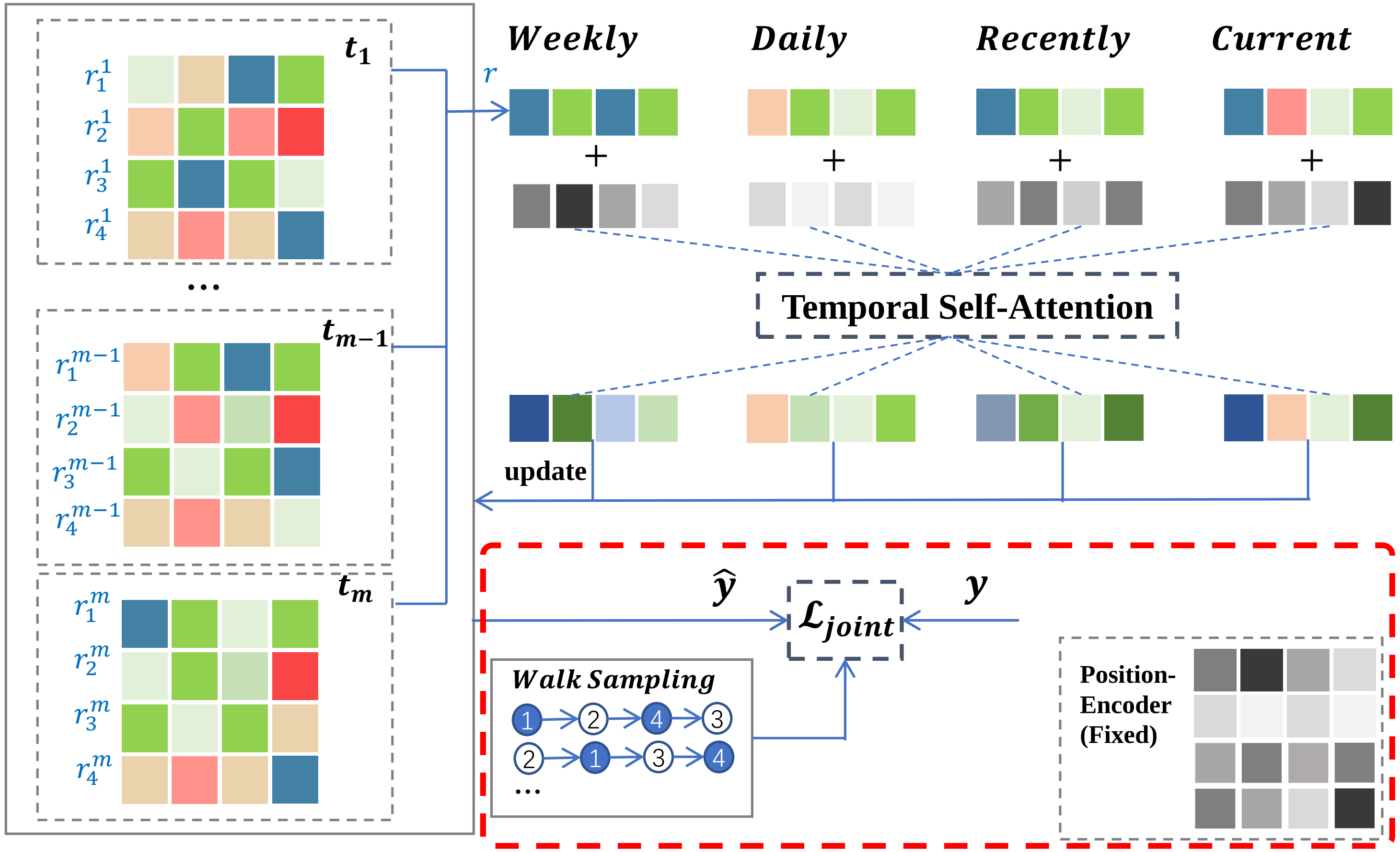
$$\mathcal{L}_{volume} = \sum_{t \in T} \sum_{r_i \in \mathcal{M}} \left| y_i^t - \frac{\sum_j^k s_{ij}^t y_j^t}{\sum_j^k s_{ij}^t} \right|$$

Final objective function

$$\mathcal{L}_{joint} = \mathcal{L}_{walk} + \mathcal{L}_{volume} + \frac{\lambda}{2} ||\mathbf{\Theta}||^2$$

Traffic volume inference

$$\hat{y}_i^t = \frac{\sum_j^k s_{ij}^t y_j^t}{\sum_j^k s_{ij}^t}$$



Outline

- Background
- Problem
- Framework
- **Experiment**
- Conclusion

Dataset

Table 1. Basic statistics of two datasets

Dataset	Hangzhou City	Jinan City
Time spans	2021/01/03-01/03	2016/08/01-08/31
# Road segments	553	493
# Monitored segments	46	165
# Features	8	7
Time interval (minute)	5	5
Sensor type	Traffic radar	Surveillance camera

Performance Study

$$RMSE = \sqrt{\frac{1}{n|T|}\sum_{t=1}^{|T|}\sum_{i=1}^{n}(y_i^t - \hat{y}_i^t)^2}$$

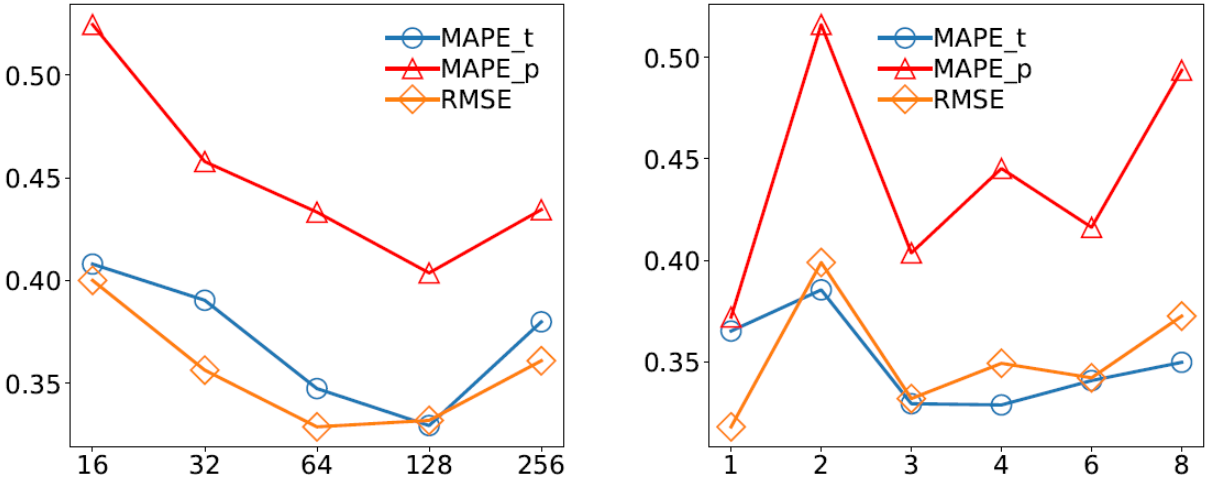
$$MAPE_t = \frac{100\%}{n|T|}\sum_{t=1}^{|T|}\sum_{i=1}^{n}\left|\frac{y_i^t - \hat{y}_i^t}{y_i^t}\right|$$

$$MAPE_p = \frac{100\%}{n|T|}\sum_{t=1}^{|T|}\sum_{i=1}^{n}\left|\frac{y_i^t - \hat{y}_i^t}{\hat{y}_i^t}\right|$$

Table 2. Performance comparison of different baselines.

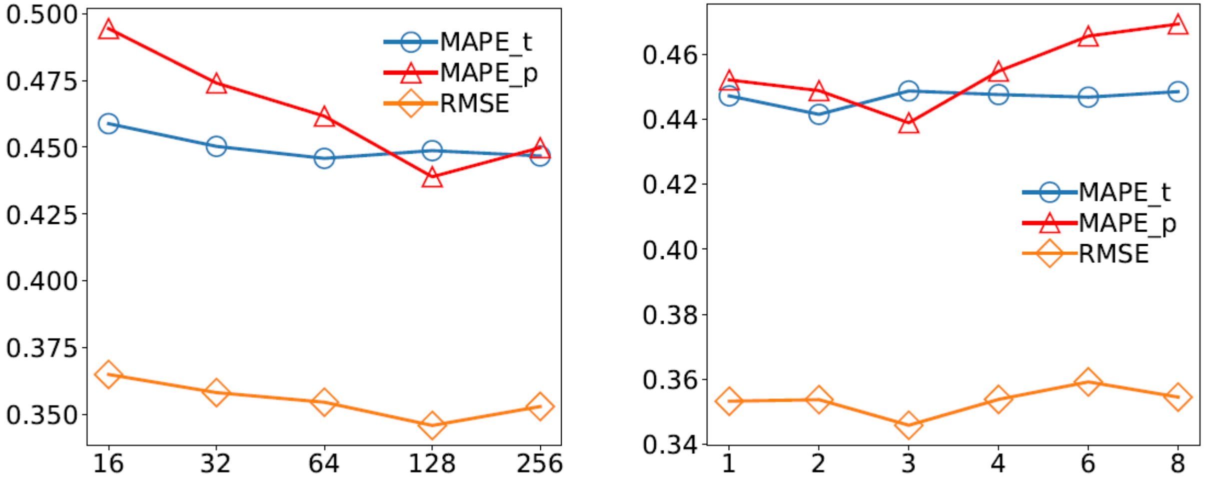
Dataset	Hangzhou City			Jinan City		
Methods	$MAPE_t$	$MAPE_p$	RMSE	$MAPE_t$	$MAPE_p$	RMSE
KNN (k=5)	0.6636	0.7139	63.1035	0.6446	0.6306	60.3842
CA (k=5)	0.6879	0.7325	65.4562	0.6568	0.6423	61.2357
MLP	0.6029	0.6561	56.4201	0.8180	0.6808	69.3974
XGBoost	0.4689	0.5243	53.9832	1.5811	0.5917	93.3649
ST-SSL	0.5638	0.5983	44.2793	0.7052	0.6883	59.0377
CT-Gen	0.3602	0.4622	37.9691	0.6727	0.4760	57.4482
JMDI	\	\	\	0.4655	0.5574	42.0020
CTVI	**0.3294**	**0.4037**	**33.1924**	**0.4487**	**0.4389**	**34.5814**

Parameter Sensitivity



(a) $w.r.t\ d$ (b) $w.r.t\ \#head$

Fig. 1. Parameter sensitivity on Hangzhou.



(a) $w.r.t\ d$ (b) $w.r.t\ \#head$

Fig. 2. Parameter sensitivity on Jinan

Outline

- Background
- Problem
- Framework
- Experiment
- **Conclusion**

Conclusion

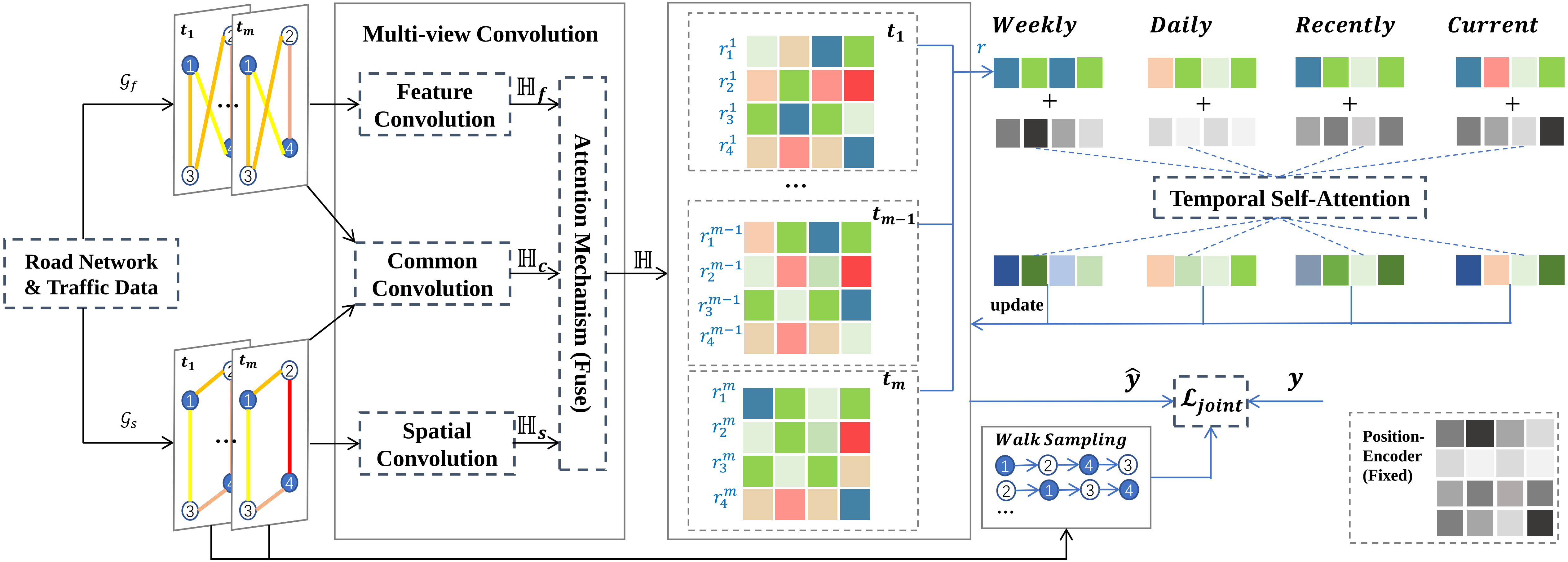
- We propose a novel framework, called CTVI, to infer citywide traffic volume by modeling complex spatial corrections and temporal dependencies.
- We incorporate **multi-view graph convolution** on spatial and feature affinity graphs with **temporal self-attention mechanism** to learn road segment representation.
- We combine an **unsupervised** random walk enhancement and a **semi-supervised** spatial-temporal volume constraint to augment the final representation.

Q&A

Thanks!

Code: https://github.com/dsj96/CTVI-master

Framework



Road Network & Traffic Data
$\mathcal{G}_f$
$\mathcal{G}_s$
t_1
t_m
Multi-view Convolution
Feature Convolution
Common Convolution
Spatial Convolution
$\mathbb{H}_f$
$\mathbb{H}_c$
$\mathbb{H}_s$
Attention Mechanism (Fuse)
$\mathbb{H}$
t_{m-1}
r_1^1 r_2^1 r_3^1 r_4^1
r_1^{m-1} r_2^{m-1} r_3^{m-1} r_4^{m-1}
r_1^m r_2^m r_3^m r_4^m
Weekly
Daily
Recently
Current
r
Temporal Self-Attention
update
$\hat{y}$
$\mathcal{L}_{joint}$
y
Walk Sampling
Position-Encoder (Fixed)