24H-GNN: A Pattern-Driven Pseudo-Label Fusion Graph Network for City-Wide Traffic Flow Prediction

Thanh Nhan Ho[1], Dong Anh Nguyen [2], Minh Tai Pham Nguyen[3], Minh Khue Phan Tran[4], Hoang Minh Nguyen[5], Nguyen Bao Tran[6], and Trong Nhan Le [1]

[1] Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam
nhan.honhanhocse@hcmut.edu.vn,trongnhanle@hcmut.edu.vn
[2] School of Industrial Management, Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam
anh.nguyendong@hcmut.edu.vn
[3] Faculty of Advanced Program, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
2151013086tai@ou.edu.vn
[4] Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
khue.ptm@ou.edu.vn
[5] Faculty of Foreign Languages, Ho Chi Minh City University of Foreign Languages – Information Technology, Ho Chi Minh City, Vietnam
22dh712862@st.huflit.edu.vn
[6] Faculty of Mechanical Engineering, Ho Chi Minh City University of Technology, Ho Chi Minh City,Vietnam
bao.trannguyen0408@hcmut.edu.vn

Abstract — **Urban traffic flow prediction is essential for effective transportation management but remains challenging due to heterogeneous and spatially inconsistent data sources. We propose a pseudo-label fusion framework that integrates loop detector data and taxi GPS trajectories using spatio-temporal similarity. Graph neural networks (GNNs) are employed to model network dependencies, while an adaptive fusion strategy balances reliable sensor data and trajectory-based estimates. The proposed approach achieves strong hotspot detection performance (*hotspot*@50 = 0.92, *hotspot*@100 = 0.97) and consistently outperforms baseline models, particularly under limited supervision. The framework enables scalable city-wide traffic estimation through multi-source data integration.**

1 Introduction

Rapid urbanization increases the demand for effective city-wide traffic management. Accurate traffic flow prediction requires large-scale data covering the entire urban network [1]. Loop detector data and taxi trajectory data provide complementary spatial coverage [2]. However, loop detectors are accurate but

sparsely deployed, whereas taxi trajectories are abundant but noisy and biased. Integrating these heterogeneous sources is challenging, particularly due to missing labels [3].

We propose a pseudo-label fusion framework to predict 24-hour traffic flow for each road segment by transferring supervision from loop-equipped segments to trajectory-only segments. All data are normalized into 24-hour patterns, and pseudo-labels are generated using spatio-temporal similarity with nearby loop detectors. The fused representations enable reconstruction of a city-wide traffic profile under sparse supervision. We evaluate the proposed approach against multiple GNN baselines.

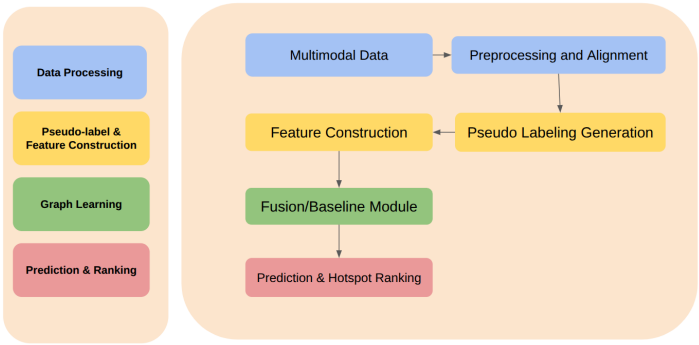


Fig. 1: Overall pipeline of the proposed 24H-GNN framework. The process includes (1) multimodal data preprocessing and spatial alignment, (2) cosine-similarity-based pseudo-label generation with spatio-temporal feature construction, and (3) fusion or baseline graph modules followed by a prediction head for traffic flow estimation and hotspot detection.

2 Related Works

Modern methods mainly use Graph Neural Networks (GNNs), including ASTGCN [4], DCRNN [5], GCN [6], GraphSAGE [7]. These models learn spatio-temporal interactions across road networks to analyze traffic. The GNN baselines serve as the main comparison set in this study for assessing traffic flow prediction performance. Overall, their development shows a growing effort to use graph-structured data for large-scale urban traffic modeling.

3 Methodology

3.1 Fusion Dataset with Pseudolabeling

To extend traffic estimation across the entire urban area, the framework combines two data sources that cover nearly the entire city but are not homogeneous

into a synthesized 24-hour pattern dataset. Since trajectory data [8] lacks labels and cannot directly provide traffic density, inference methods are needed to address this issue. Pseudo-labels are constructed at the road-segment level for taxi trajectory data that have been map-matched onto the city's road network. These pseudo-labels are then calibrated using data from loop detectors at sensor locations. The process includes 24-hour pattern normalization, pseudo-label assignment based on cosine similarity, merging of taxi and pseudo-labeled data, and creating a complete dataset across the entire network.

For each segment without a loop detector, the five loop-equipped segments $N_K(j)$ exhibiting the most similar 24-hour traffic patterns s_i. y_i^{loop}denotes the normalized 24-hour flow pattern recorded by loop detector i, and p_j^{taxi} represents the trajectory-derived temporal profile of segment j, K is set to 5 to balance local similarity and data sparsity..

$$s_i = \cos\left(y_i^{\text{loop}}, p_j^{\text{taxi}}\right), \quad i \in L$$

$$N_K(j) = 5$$

Each weighted average thus becomes a pseudo-label, a reliable estimate of the traffic flow that a given road segment would exhibit if it were equipped with a sensor. In this way, biased or incomplete data are transformed into reference-quality data suitable for machine learning models, leveraging the inherent stability of traffic patterns.

$$w_i = \frac{s_i}{\sum_{k \in \mathcal{N}_K(j)} s_k} \tag{1}$$

$$\mathbf{y}_j^{\text{pseudo}} = \sum_{i \in \mathcal{N}_K(j)} w_i \, \mathbf{y}_i^{\text{loop}} \tag{2}$$

To focus on stable long-term patterns, irregular days (e.g., holidays or special events) are excluded so that pseudo-labels reflect normal traffic conditions.

Merging Trajectory Dataset: After pseudo-labels y_j^{pseudo} are assigned to all 838 road segments, each road segment is represented by two components: (1) a relative traffic flow vector p_j^{taxi} and (2) a pseudo-label y_j^{pseudo} derived from reliable sensor-based data. These two sources are then combined through an adjusted fusion operation as follows:

$$\tilde{y}_j = \alpha \cdot y_j^{\text{pseudo}} + (1 - \alpha) \cdot p_j^{\text{taxi}}, \quad \alpha = 0.7$$

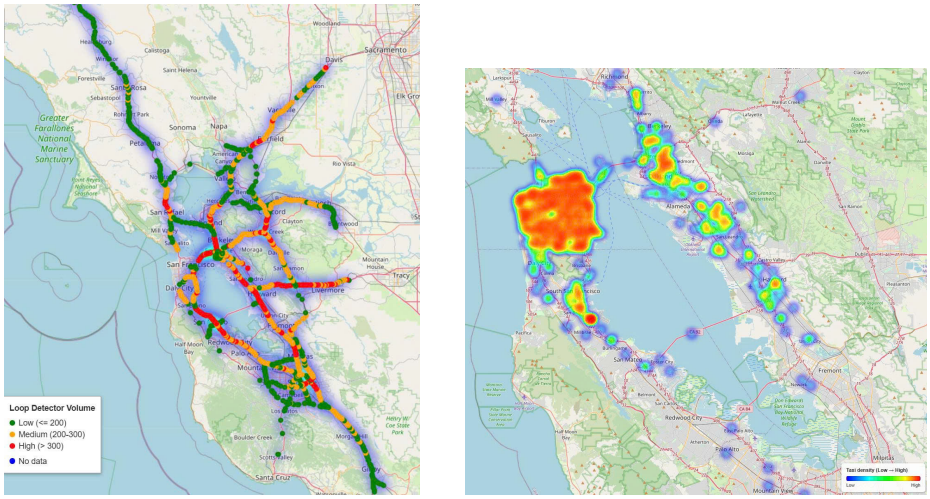
$\alpha = 0.7$ is an empirically chosen coefficient within the [0.6-0.8] range that ensures the model preserves the stability provided by the pseudo-labels while still reflecting real-world variations and local deviations observed in the taxi data.

4 Experiment

4.1 Dataset

For loop detector data, we use the LagerST subset [9] which is used for large-scale traffic forecasting in Greater Bay Area (GBA), it covers 2352 sensors in 2020 and

2021. Trajectory dataset with GPS Taxi is the SFMTA Taxi Trips, which records from all permitted taxi companies in the City and County of San Francisco.



(a) Loop detectors location. (b) Trajectory taxi trips heatmap.

Fig. 2: Loop Detectors And Trajectory Taxi Dataset.

4.2 Results

Table 1: Training and Hyperparameter Settings

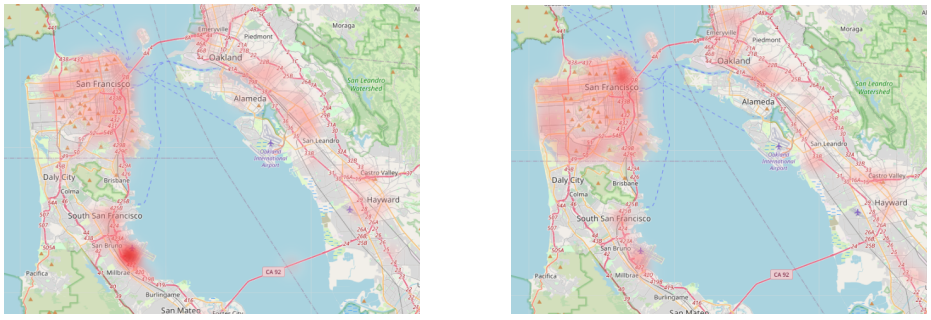
Parameter	**Value**
Optimizer	Adam
Learning rate	1.0×10^{-4}
Number of neighbors (K)	5
Fusion coefficient (α)	0.7
Epochs	50
Batch size	10

Method	R^2	*hotspot*@50	*hotspot*@100
ASTGCN	0.71	0.6	0.75
DCRNN	0.74	0.64	0.74
STGCN	0.76	0.68	0.74
GCN	0.75	0.62	0.7
GraphSAGE	**0.82**	0.84	0.93
STGFormer	0.51	0	0.22
Fusion-GraphSAGE(Ours)	0.81	0.88	**0.97**
Fusion-MultiScaleGNN(Ours)	0.79	**0.92**	0.93

Table 2: Model Performance Metrics Across Different Models

Table 2 summarizes the performance of all baseline models and the proposed fusion-based frameworks. Overall, the proposed methods consistently outperform the baselines, confirming the effectiveness of adaptive multi-source fusion. Fusion-MultiScaleGNN achieves the highest hotspot detection accuracy at a smaller ranking threshold (*hotspot*@50 = 0.92), while Fusion-GraphSAGE performs best at a larger ranking scale (*hotspot*@100 = 0.97), indicating complementary strengths in local and global congestion identification. GraphSAGE achieves strong $\mathbf{R^2}$ performance.

4.3 Hotspot Visualization



(a) Ground Truth Top 100 Hotspots. (b) Prediction of Top 100 Hotspots.

Fig. 3: Results with top 100 hotspots detection at 16-17 hours

The Figure 3, illustrates the model's ability to detect congestion hotspots by comparing predicted results with ground-truth data. Figure visualizes the spatial distribution of the top 100 hotspots in the San Francisco Bay Area during the peak period (16–17 hours).

Across both maps, the visualization shows that the predicted hotspot locations are quite close to the actual congestion patterns, identifying several key

bottlenecks in downtown San Francisco and key arterial routes, an essential capability for practical urban traffic management applications.

5 Conclusion

This study proposes a pseudo-label fusion framework for city-wide traffic prediction. By integrating loop detector and trajectory data, the model effectively learns 24-hour traffic patterns and improves hotspot detection. The spatio-temporal pseudo-labeling mechanism extends supervision to unlabeled regions, enabling scalable traffic estimation under limited ground-truth data.

6 Acknowledgements

This work is acknowledged by Ho Chi Minh City University of Technology - Vietnam National University and Ho Chi Minh City Open University, Vietnam.

References

1. Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2015.
2. C. Meng, X. Yi, L. Su, J. Gao, and Y. Zheng, "City-wide traffic volume inference with loop detector data and taxi trajectories," SIGSPATIAL '17, (New York, NY, USA), Association for Computing Machinery, 2017.
3. S. Li and H. Tang, "Multimodal alignment and fusion: A survey," 2025.
4. J. Zhu, Q. Wang, C. Tao, H. Deng, L. Zhao, and H. Li, "Ast-gcn: Attribute-augmented spatiotemporal graph convolutional network for traffic forecasting," *IEEE Access*, vol. 9, pp. 35973–35983, 2021.
5. Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," 2018.
6. T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2017.
7. W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," 2018.
8. Y. Zheng, "Trajectory data mining: An overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, May 2015.
9. X. Liu, Y. Xia, Y. Liang, J. Hu, Y. Wang, L. Bai, C. Huang, Z. Liu, B. Hooi, and R. Zimmermann, "Largest: A benchmark dataset for large-scale traffic forecasting," 2023.