NEURAL INFORMATION PROCESSING SYSTEMS

SynMob: Creating High-Fidelity Synthetic GPS Trajectory Dataset for Urban Mobility Analysis

Yuanshao Zhu[1,2*], **Yongchao Ye**[1*], **Ying Wu**[1,3], **Xiangyu Zhao**[2†], **James J.Q. Yu**[4†]

[1] Southern University of Science and Technology

[2] City University of Hong Kong

[3] University of Leeds

[4] University of York

{zhuys2019, 12032868, 12059004}@mail.sustech.edu.cn

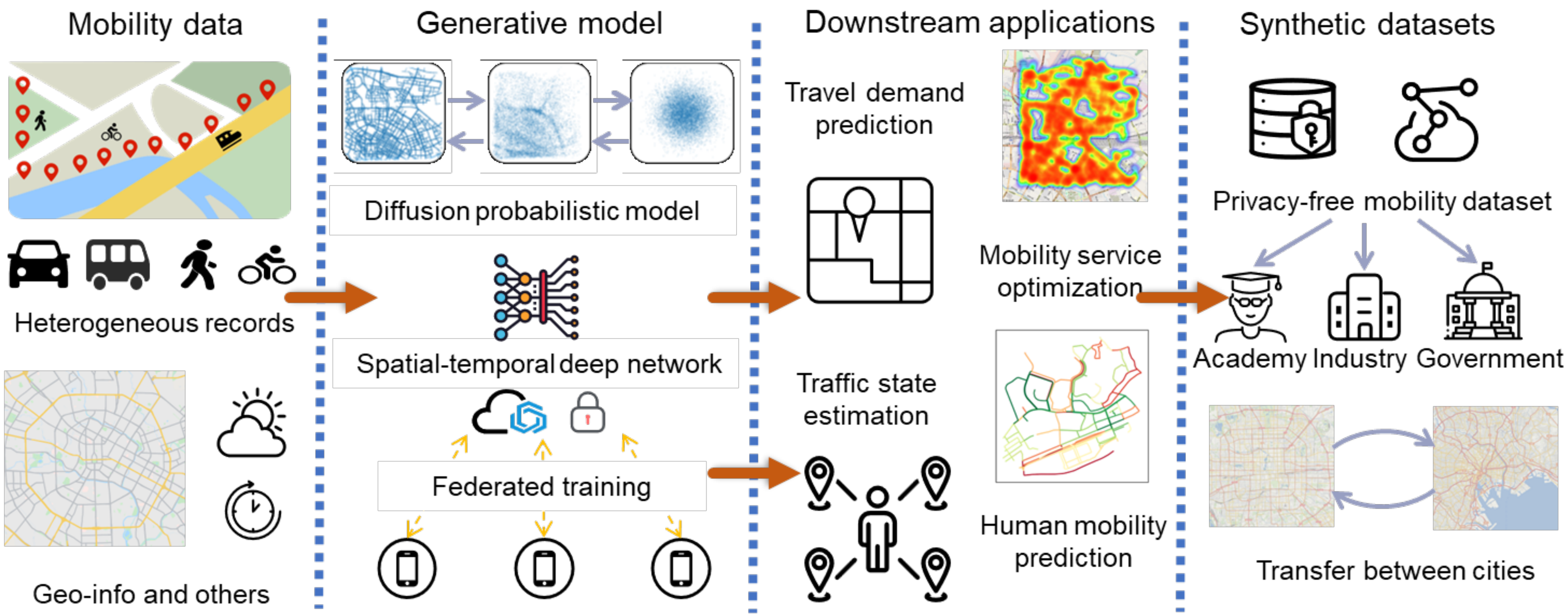
xianzhao@cityu.edu.hk

james.yu@york.ac.uk

Motivations

◆ Urban mobility analysis

- Data sources
  - GPS
  - Public transportation records
  - Application usage
- Applications
  - Transportation planning and management
  - Public health and epidemiology

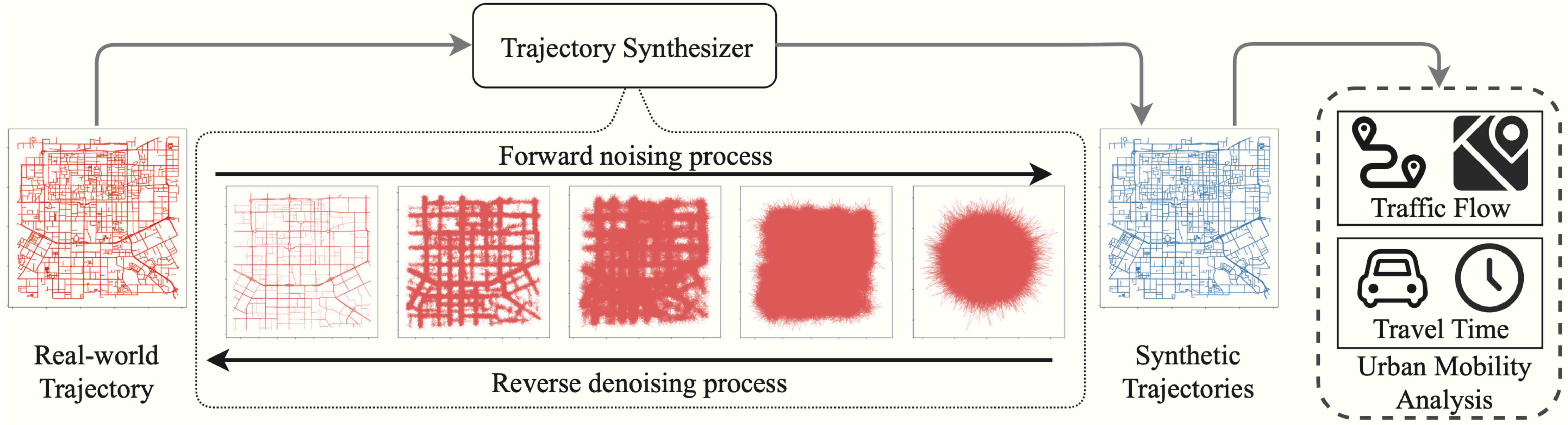


Existing Dataset Issues

- Lack of publicly available trajectory datasets.
- Strict regulations and data privacy concerns limit the accessibility of trajectory data.
- Suffering from inconsistent format and poor quality.

Dataset	GPS trajectory	Availability	Data quality	Privacy	# Trajectory
GeoLife[3] [47]	✓	✓	✗	✗	17,621
T-drive[4] [39]	✓	✓	✗	✗	10,357
Porto[5] [22]	✓	✓	✗	✗	1.7 million
Foursquare[6] [37]	✗	✓	–	✗	104,478
NYC[7]	✗	✓	–	✗	1.1 billion
Taxi-Shanghai[8]	✓	✗	✗	✗	1.2 million
GAIA[9]	✓	✗	✓	✗	3.1 million
Ours (Synthetic)	✓	✓	✓	✓	unrestricted (customize)

Technical Design



Using diffusion model as the trajectory synthesizer

Dataset Analysis

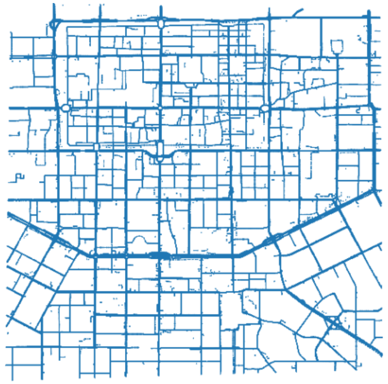
- Original dataset

Ride-hailing trajectories in Chengdu & Xi’an

Dataset	Trajectory Number	Average Time	Average Distance
Chengdu	3 493 918	11.42 min	7.42 km
Xi’an	2 180 348	12.58 min	5.73 km

Chengdu Trajectory

Chengdu Heatmap



Xi’an Trajectory

Xi’an Heatmap

Dataset Analysis

1. Overview of two synthetic dataset

- ✓ **Privacy free**: It provides privacy protection by generating trajectories process.
- ✓ **High fidelity**: It has high fidelity, with similar statistical features as the original dataset.
- ✓ **Public availability**: It publicly available without violating regulations.
- ✓ **Scalability**: It can generate an arbitrary amount of synthetic trajectories.
- ✓ **Enhancing diversity**: It offers various trajectory patterns.

Table 2: Dataset description of SYN-CHENGDU

Type	Description
Format	pickle / geoparquet
Size	4.39 GB
Value type	float64
Time frame	5 min
Sample interval	3 s
Spatial coverage	lat: $30.65° \sim 30.73°$ lng: $104.04° \sim 104.13°$

Table 5: Dataset description of SYN-XI'AN

Type	Description
Format	pickle / geoparquet
Size	4.66 GB
Value type	float64
Time frame	5 min
Sample interval	3 s
Spatial coverage	lat: $34.20° \sim 34.28°$ lng: $108.90° \sim 108.99°$

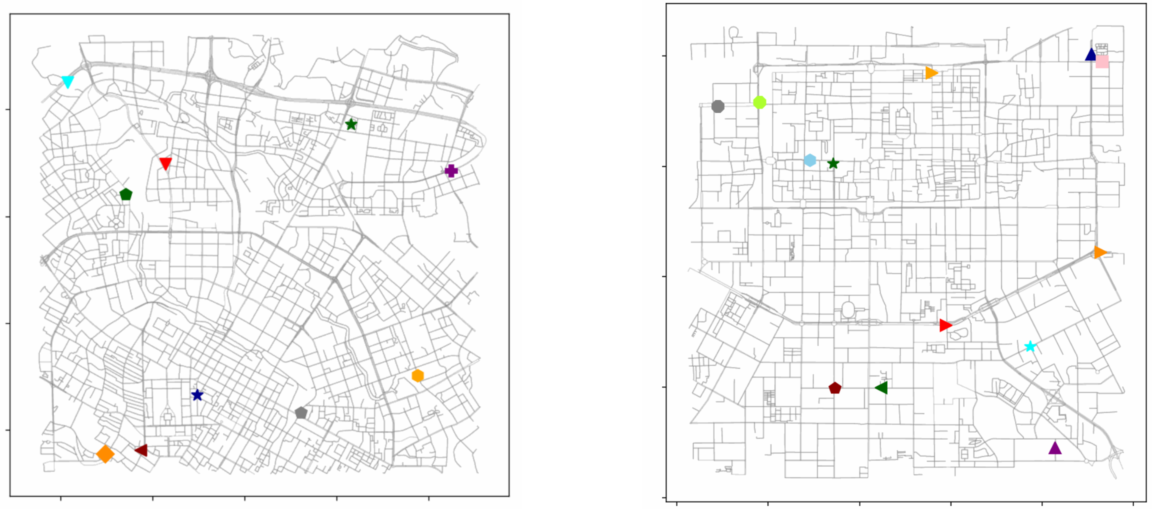
Dataset Analysis

2. Geographic visualization (cases)

The synthetic trajectories are richly diverse, reflecting multiple and complex mobility patterns.

Same origin and destination

The synthetic trajectories are able to cover the entire urban area and follow the path of the road network.



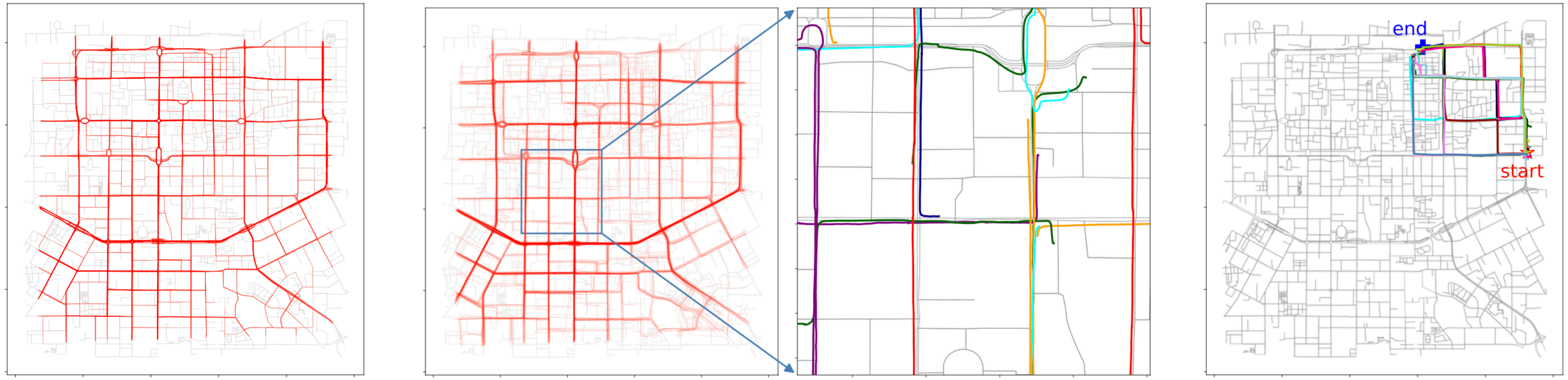
Different origin and destination

Dataset Analysis

3. Trajectories geo-distribution insight

The synthetic trajectories successfully adhere to the geo-distribution and sparse properties compare to the original counterparts.It can also maintain consistent start and end areas.

(a) Original trajectories. (b) Synthetic trajectories with area zoom. (c) Same start-end areas.

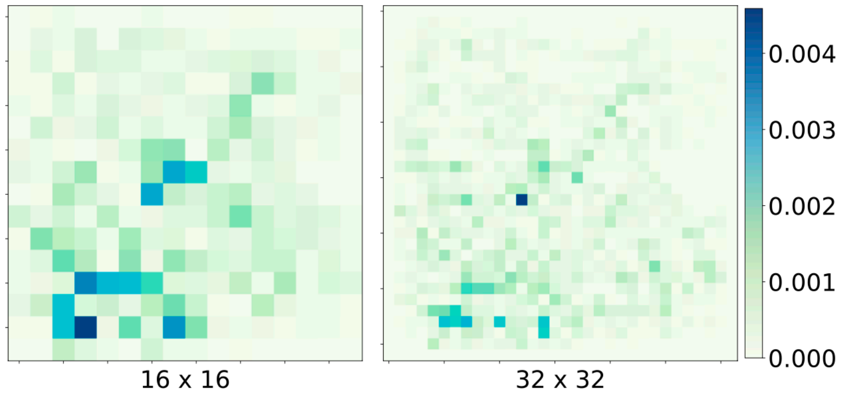


(a) Original trajectories. (b) Synthetic trajectories with area zoom. (c) Same start-end areas.

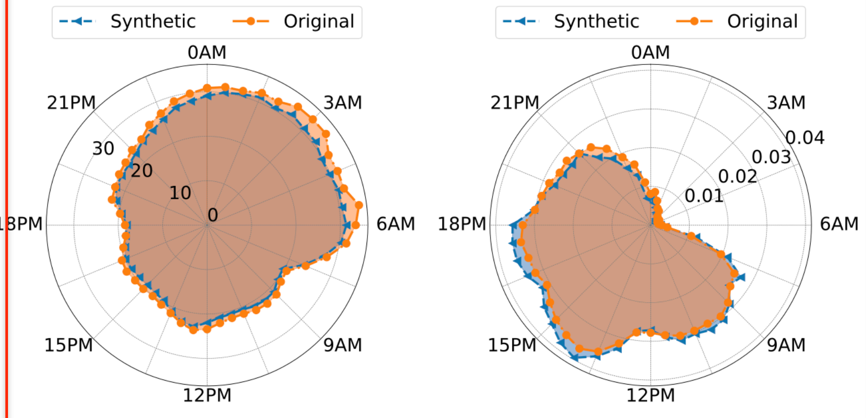
Dataset Analysis

4. Spatial-temporal distribution

- The synthetic trajectory dataset can maintain a high degree of spatial distribution consistency.

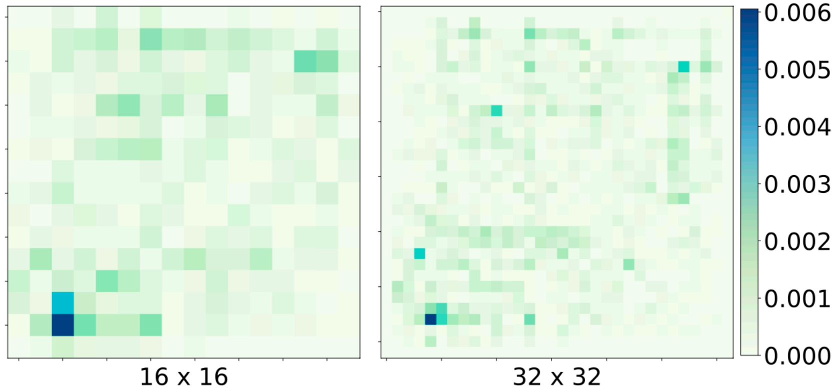


(a) Heatmap distribution of difference

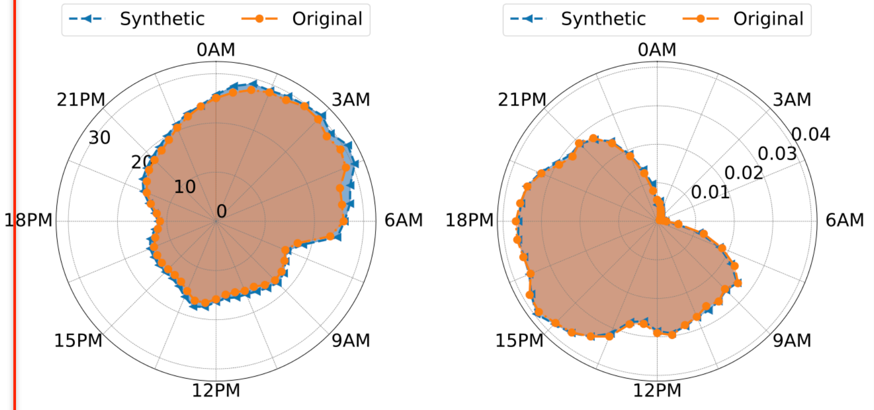


(b) Average speed (c) Trip volume

- The synthetic trajectory dataset also ensures consistent distribution at the temporal level.



(a) Heatmap distribution of difference

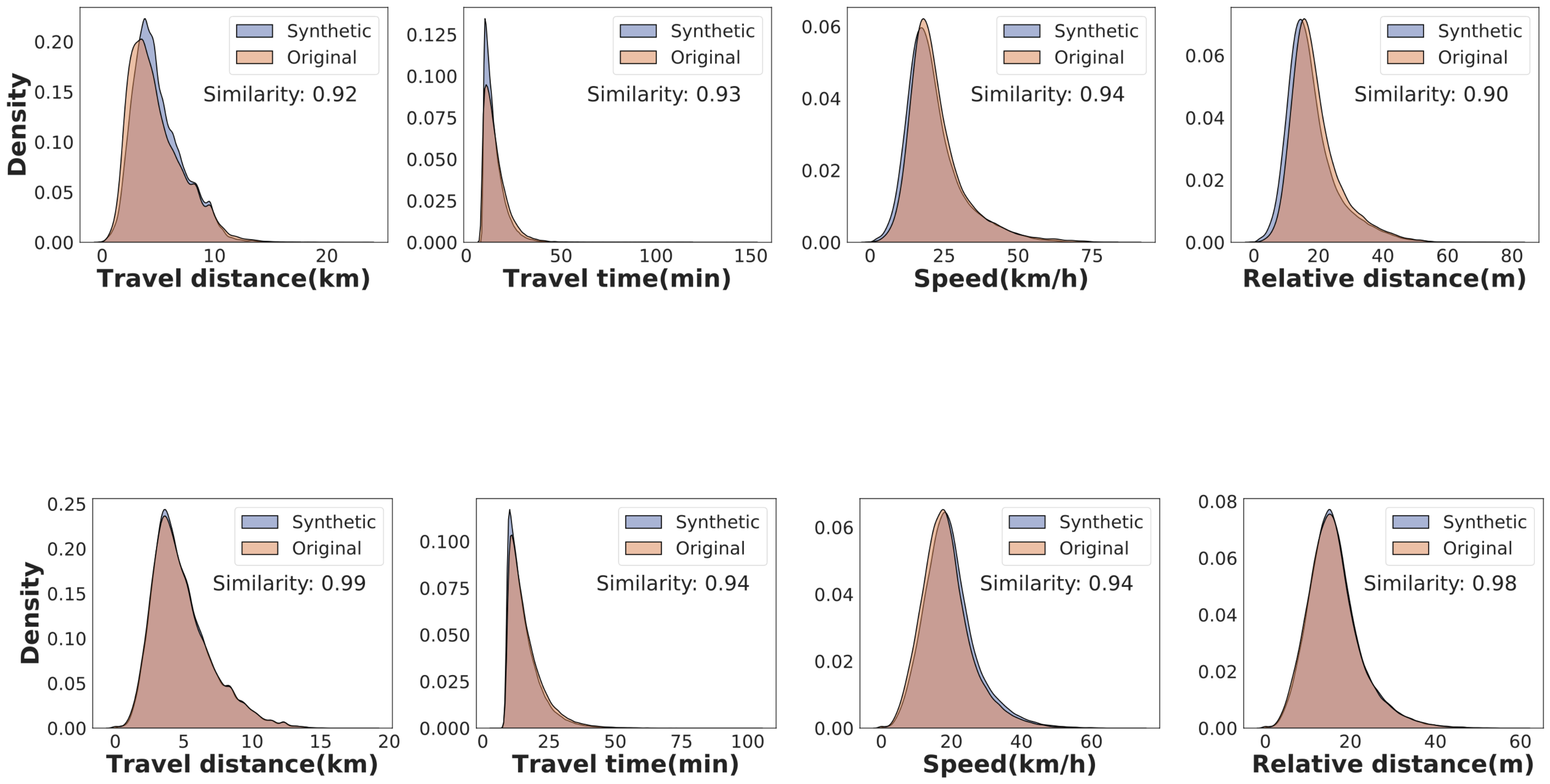


(b) Average speed (c) Trip volume

Dataset Analysis

5. Trajectories properties

The synthetic dataset showing a strong adherence to the trajectory level properties observed in the original data



Dataset Analysis

6. Data utility case studies

Travel demand prediction

Problem: predict the vehicle inflow or outflow x_d^t for a given area d at time t (time-series prediction)

$$[x^{t-H+1}, \ldots, x^t] \rightarrow [x^{t+1}, \ldots, x^{t+h}]$$

Methods: representative ST prediction model

original / synthetic / difference ratio (%)

Methods	AGCRN	GWNet	DCRNN	MTGNN
RMSE	6.91 / 6.50 / 5.93%	6.90 / 6.53 / 5.36%	7.29 / 6.48 / 11.11%	6.81 / 6.41 / 5.87%
MAE	4.64 / 4.43 / 4.53%	4.65 / 4.47 / 3.87%	4.88 / 4.45 / 8.81%	4.58 / 4.39 / 4.15%
MAPE	30.47/ 30.97 / 1.64%	30.57 / 30.74 / 0.56%	32.40 / 30.40 / 6.17%	29.61 / 29.88 / 0.91%

Travel time estimation

Problem: estimate the travel time between a pair of origins and destinations

$$[o, d, t, V] \rightarrow y, V = \{v_1, v_2, \ldots, v_m\}$$

Methods: representative ML/DL TTE models

original / synthetic / difference ratio (%)

Methods	TEMP	XGBoost	WDR	DeepTTE
RMSE	290.32 / 282.33 / 2.75%	271.56 / 256.19 / 5.66%	258.64 / 247.29 / 4.39%	216.93 / 193.34 / 10.87%
MAE	182.74 / 174.42 / 4.55%	175.20 / 167.75 / 4.25%	149.81 / 140.05 / 6.51%	132.95 / 121.31 / 8.76%
MAPE	18.62 / 17.81 / 4.30%	17.97 / 16.94 / 5.73%	14.06 / 13.17 / 6.33%	13.07 / 12.29 / 5.97%

Conclusion

- Conclusion
  - A trajectory generation model based on diffusion model
  - A high-fidelity synthetic trajectory dataset for urban mobility analysis

- Limitations
  - Raw data is still needed in the training stage
  - Only focus on one travel mode (ride-hailing trajectories)
  - Large computational cost of generative models

NEURAL INFORMATION PROCESSING SYSTEMS

Thank You!

Yuanshao Zhu

zhuys2019@mail.sustech.edu.cn

Southern University of Science and Technology

City University of Hong Kong