Kinematic Temporal VAE for Generalized Pedestrian Prediction

Dongchen LI, Zhimao LIN, Jinglu HU

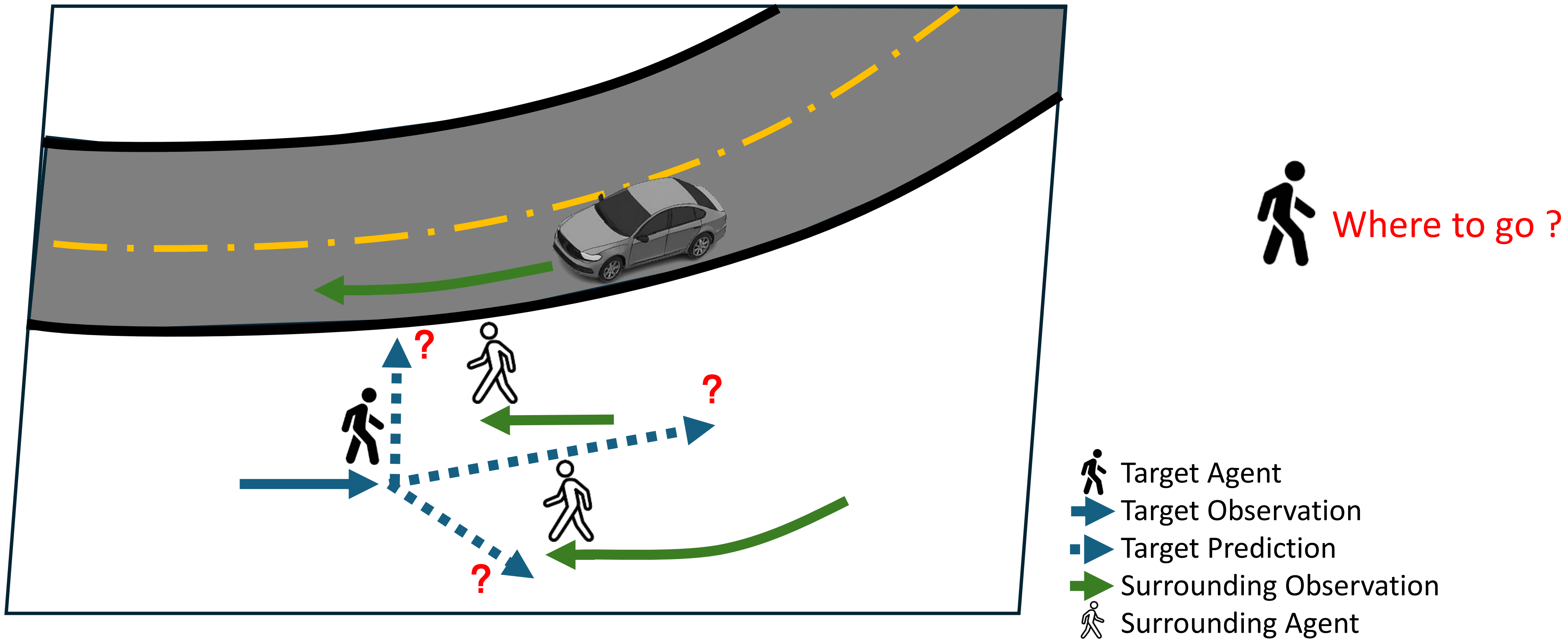
Waseda University

Graduate School of Information, Production and Systems

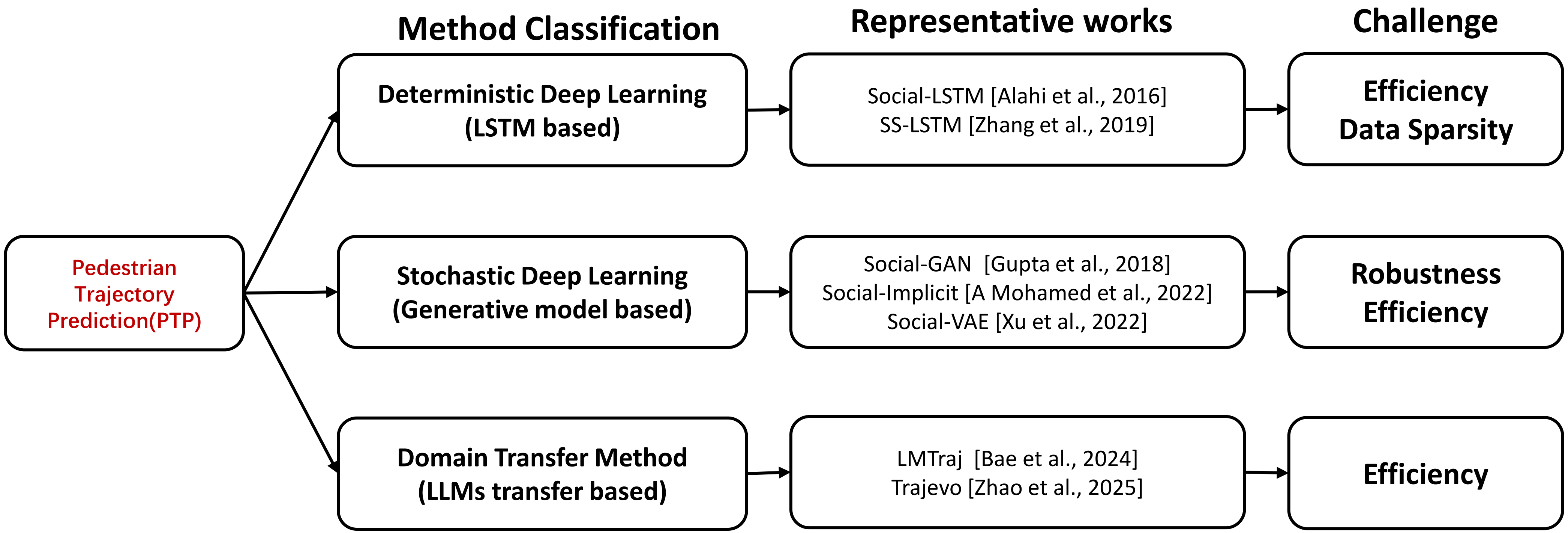
08.10.2025

Introduction

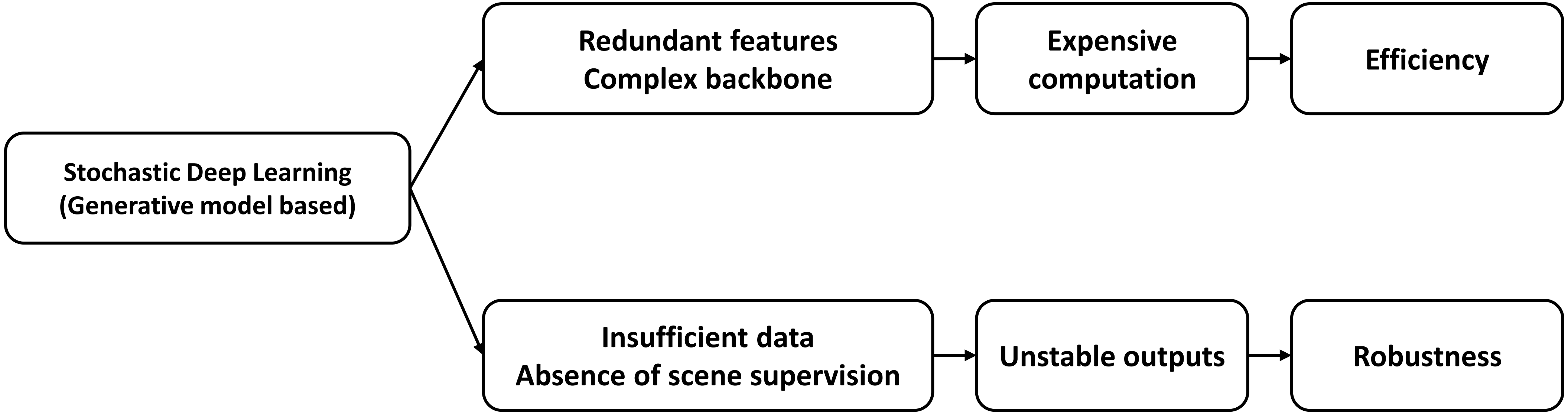
The pedestrian trajectory prediction is a crucial research topic in artificial intelligence application scenarios like autopilot and robotics.



Related Work



Detailed Challenge



Our Motivation

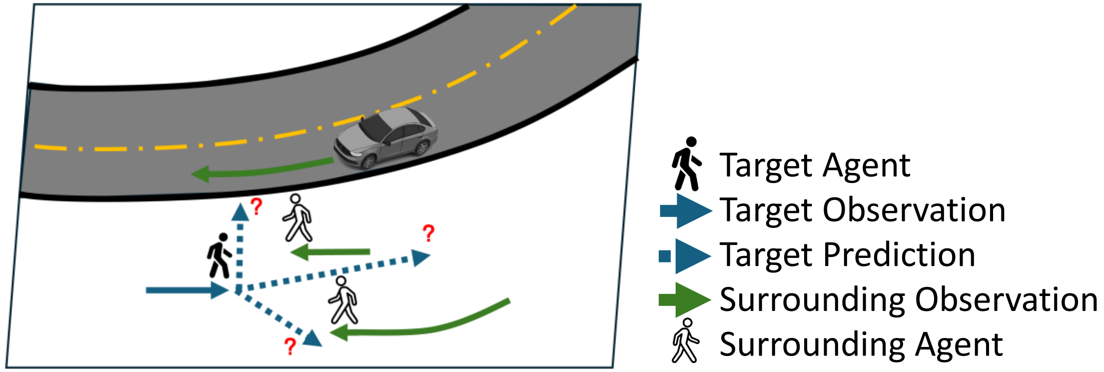
To address the robustness and efficiency issues of generative model outputs, we focus on improving the efficiency of the backbone and introducing robust post-processing.

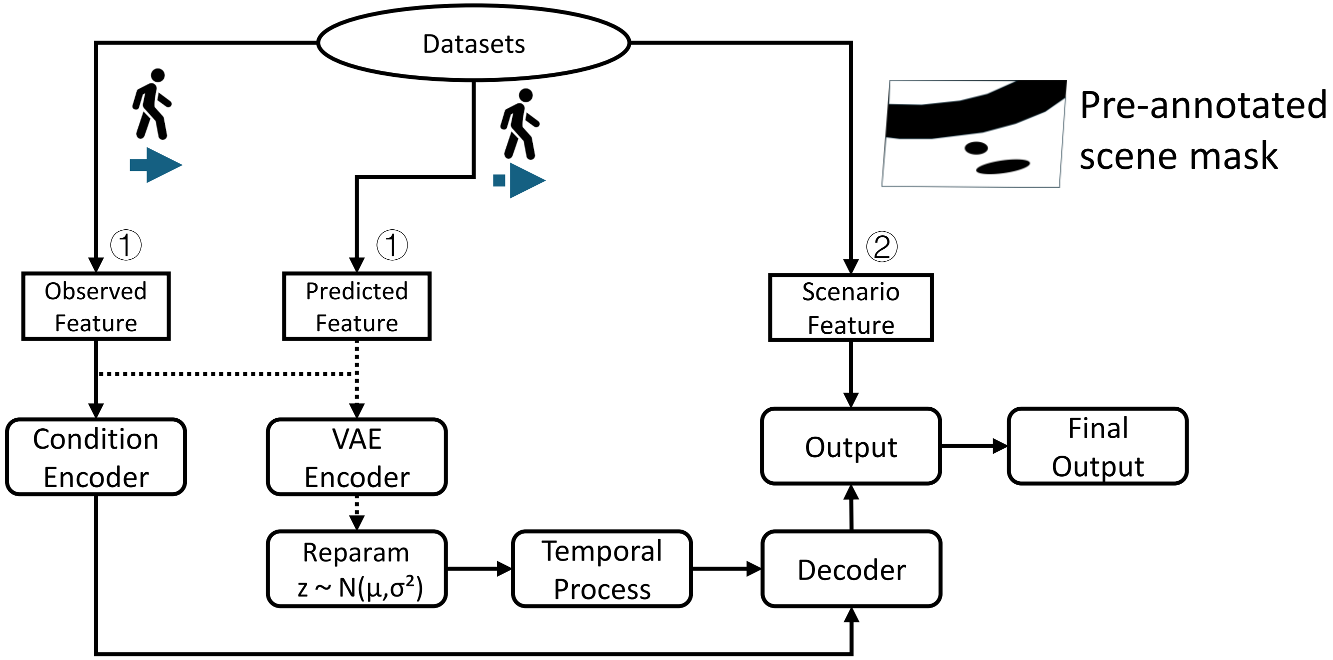
- First, we refined the model’s input representation by reducing redundant features as much as possible.
- Second, we introduced pre-annotated scene information to provide supervision, thereby enhancing the plausibility of the generated outputs.

Our Framework

① The target input is streamlined, and surrounding pedestrian information is refined to reduce redundancy and improve efficiency.

② Pre-annotated scene information is used to constrain future trajectory outputs and reduce output instability.





Experiments

Datasets：

ETH&UCY are widely used pedestrian trajectory prediction benchmarks, consisting of real-world crowd scenarios with diverse interaction patterns.

Metrics：

ADE (Average Displacement Error) – Mean Euclidean distance between predicted and ground-truth trajectories over all predicted steps (lower = better accuracy).

FDE (Final Displacement Error) – Euclidean distance between the predicted final position and the ground-truth final position (lower = better accuracy).

Inference Time – Average time required to produce a prediction for one instance (lower = faster inference).

Standard Deviation (STD) – Variation or dispersion of prediction errors, indicating generalization ability (lower = better generalization).

Experiments

TABLE I
THE PRACTICAL EXPERIMENT WITH ADE/FDE (METER)

Method	Year	Method	ETH	HOTEL	STUDENT	ZARA01	ZARA02	AVG	STD	MAX_DEV
Constant-Velocity	-	KB	0.61/1.32	0.65/1.40	0.59/1.25	0.68/1.45	0.66/1.40	0.64/1.36	0.03/0.07^{+}	0.05/0.11^{+}
Least-Squares	-	KB	0.80/1.54	0.86/1.67	0.72/1.41	0.76/1.66	0.83/1.60	0.81/1.58	0.05/0.09^{+}	0.09/0.16^{+}
Social-LSTM[24]	2017	DM	0.50/1.07	0.11/0.23	0.27/0.48	0.22/0.60	0.24/0.77	0.27/0.63^{+}	0.13/0.28	0.23/0.44
S-GAN[25]	2017	GM	0.61/0.81	0.48/0.72	0.36/0.60	0.21/0.34	0.27/0.42	0.39/0.58	0.14/0.18	0.22/0.24
TPNMS[26]	2018	DM	0.52/0.89	0.22/0.39	0.55/1.13	0.35/0.70	0.27/0.56	0.38/0.73	0.13/0.26	0.17/0.40
Social-STGCNN[6]	2020	GM	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48	0.44/0.75	0.12/0.23	0.20/0.36
Social-Implicit[27]	2022	GM	0.66/1.44	0.20/0.36	0.31/0.60	0.25/0.50	0.22/0.43	0.33/0.67	0.17/0.40	0.33/0.77
Social-VAE[2]	2022	GM	0.41/0.58	0.13/0.19	0.21/0.36	0.17/0.29	0.13/0.22	0.21/0.33^{+}	0.10/0.14	0.20/0.25
Bo-sampler[28]	2023	GM	0.52/0.95	0.19/0.39	0.30/0.67	0.14/0.33	0.20/0.45	0.27/0.56^{+}	0.14/0.22	0.25/0.39
KT-VAE	2025	GM	0.48/0.75	0.27/0.54	0.46/0.71	0.28/0.65	0.31/0.53	0.36/0.66	0.09/0.09^{+}	0.12/0.11^{+}
KT-VAE-P	2025	GM	0.44/0.73	0.26/0.54	0.44/0.65	0.23/0.60	0.26/0.50	0.33/0.61	0.11/0.09	0.12/0.13

Our proposed approach maintains a competitive level of performance. Notably, the approach maintains strong generalization across diverse scenes.

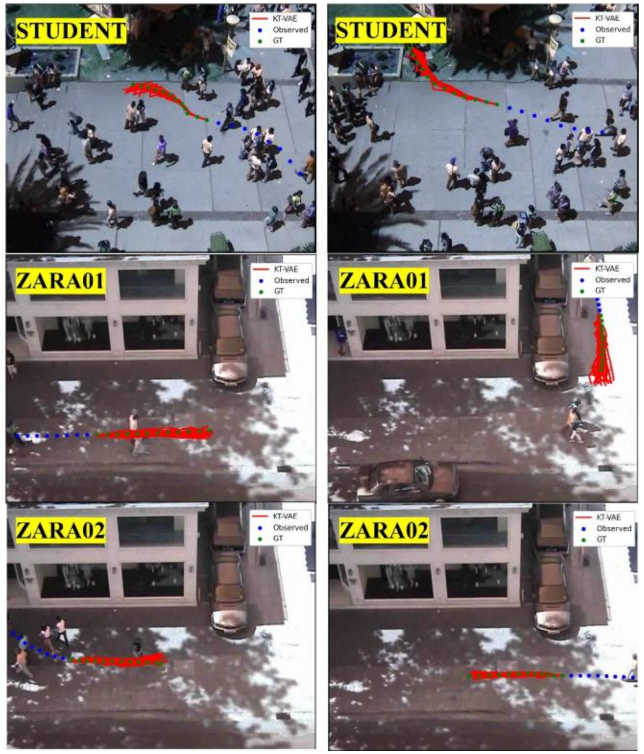
* All metrics are the smaller the better. The + means the top-3. The table above uses the average displacement error & final displacement error (Euclidean distance) metric for evaluation.

Experiments

TABLE II

THE PRACTICAL INFERENCE EXPERIMENT WITH TIME (SECOND)

Method	Time(H)	Time(L)
Constant-Velocity	0.0009	**0.01**
Least-Squares	0.0011	**0.01**
Social-LSTM[24]	0.0254	0.47
S-GAN[25]	0.0410	0.52
TPNMS[26]	0.0335	0.44
Social-STGCNN[6]	0.0175	0.18
Social-Implicit[27]	0.0087	**0.08**
Social-VAE[2]	0.4519	2.25
Bo-sampler[28]	0.0195	0.11
KT-VAE	0.0109	**0.09**
KT-VAE-P	0.0158	0.14



The table on the left presents the efficiency experiments, where the model maintains high inference efficiency. The figure on the right presents our quality analysis.

* All metrics are the smaller the better. The bolded values indicate the best performance. H: high performance device, L: Low performance device.

Conclusion

We propose KT-VAE with post-processing to reduce scenario overfitting while maintaining predictive accuracy, improving robustness and stability. Its lightweight design enables deployment on low-performance devices, meeting real-world requirements, while offering a novel spatial-temporal feature processing perspective for pedestrian trajectory prediction. In future work, we will evaluate KT-VAE under drastic spatial changes and traffic-density shifts to further enhance robustness.

Thanks for your listening