Enhancing Deep Generative Models for Distinguishing Sampling Zeros and Structural Zeros in Population Synthesis

MATSim User Meeting 2023

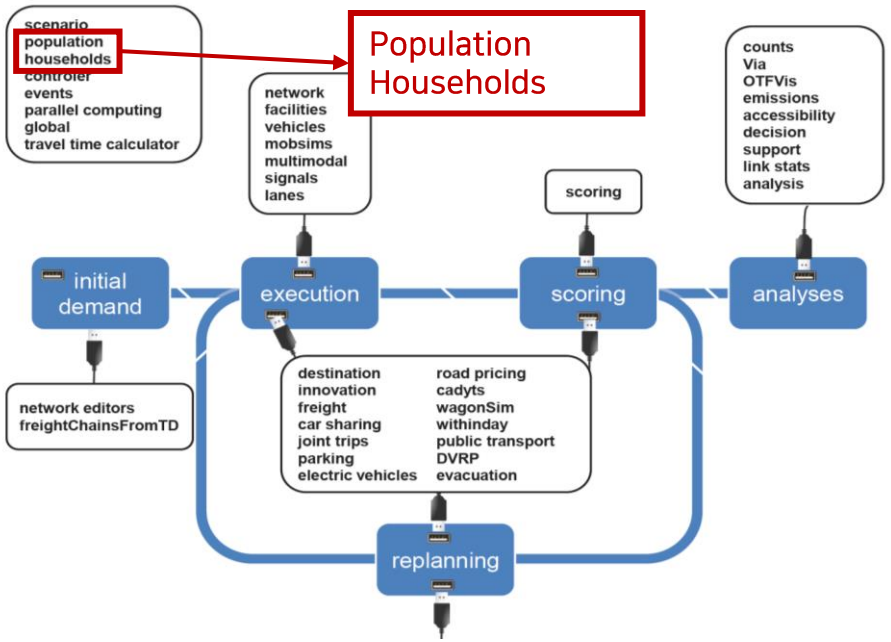
Cho Chun Shik Graduate School of Mobility
Donghyun Kwon, Inhi Kim

Introduction

Population synthesis for activity-based model (ABM)

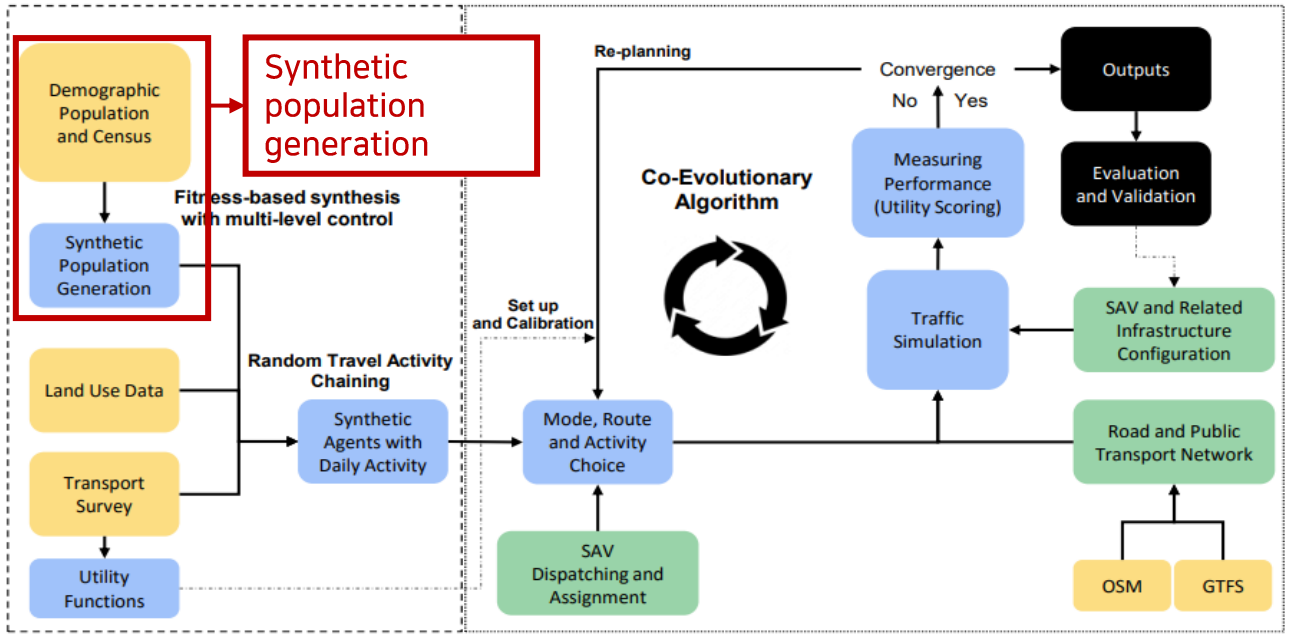
- The ABM requires a synthetic population, which serves as the essential foundation for the entire procedure.
- Our work is primarily focused on the precise generation of a synthetic population since it precedes activity generation and scheduling, influencing significantly through the subsequent modeling stages.

MATSim modeling Framework



W Axhausen, Kay, Andreas Horni, and Kai Nagel. The multi-agent transport simulation MATSim. Ubiquity Press, 2016.

Shared Autonomous Vehicle modeling Framework

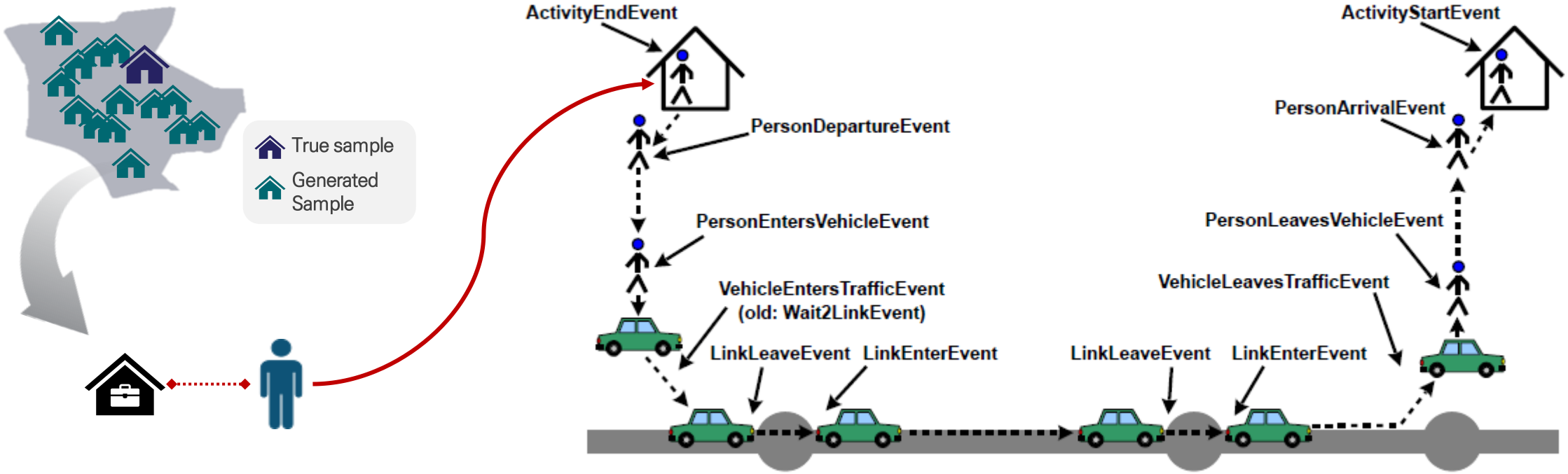


Vosooghi, Reza. Shared autonomous vehicle service design, modeling, and simulation. Diss. Université Paris Saclay (COmUE), 2019.

Introduction

Population synthesis for activity-based model (ABM)

- The synthesized population best represents the characteristics of the real population in the study region.
- In the typical modeling procedure, the activity type/location, mode choice, and trips/tours/schedules are defined and assigned based on each individual attribute information.

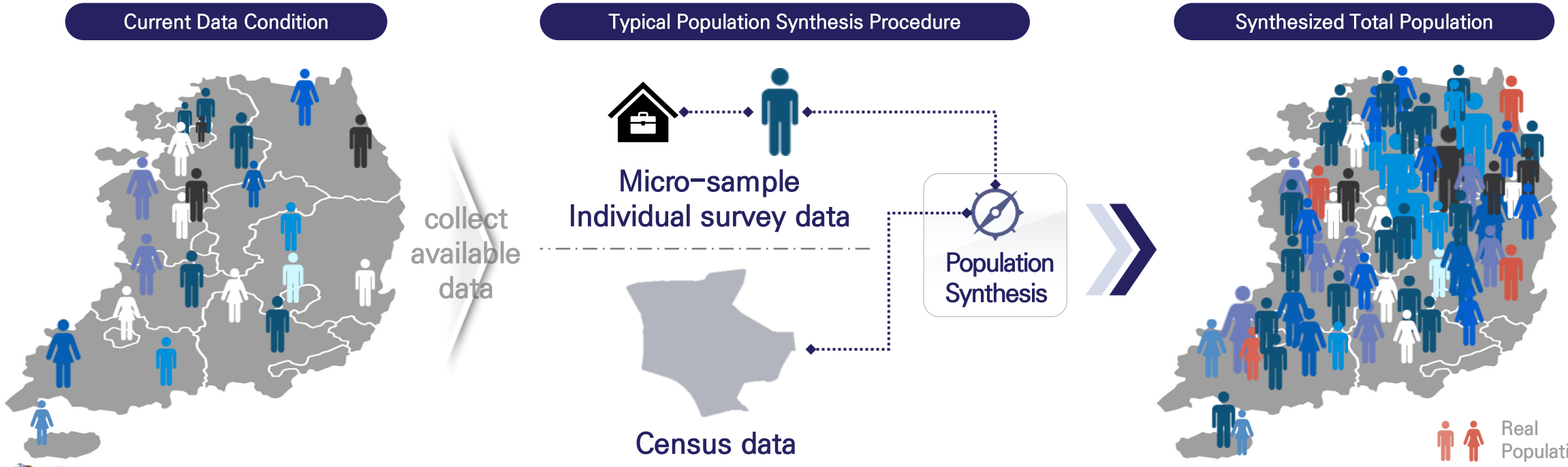


W Axhausen, Kay, Andreas Horni, and Kai Nagel. The multi-agent transport simulation MATSim. Ubiquity Press, 2016.

Related Works

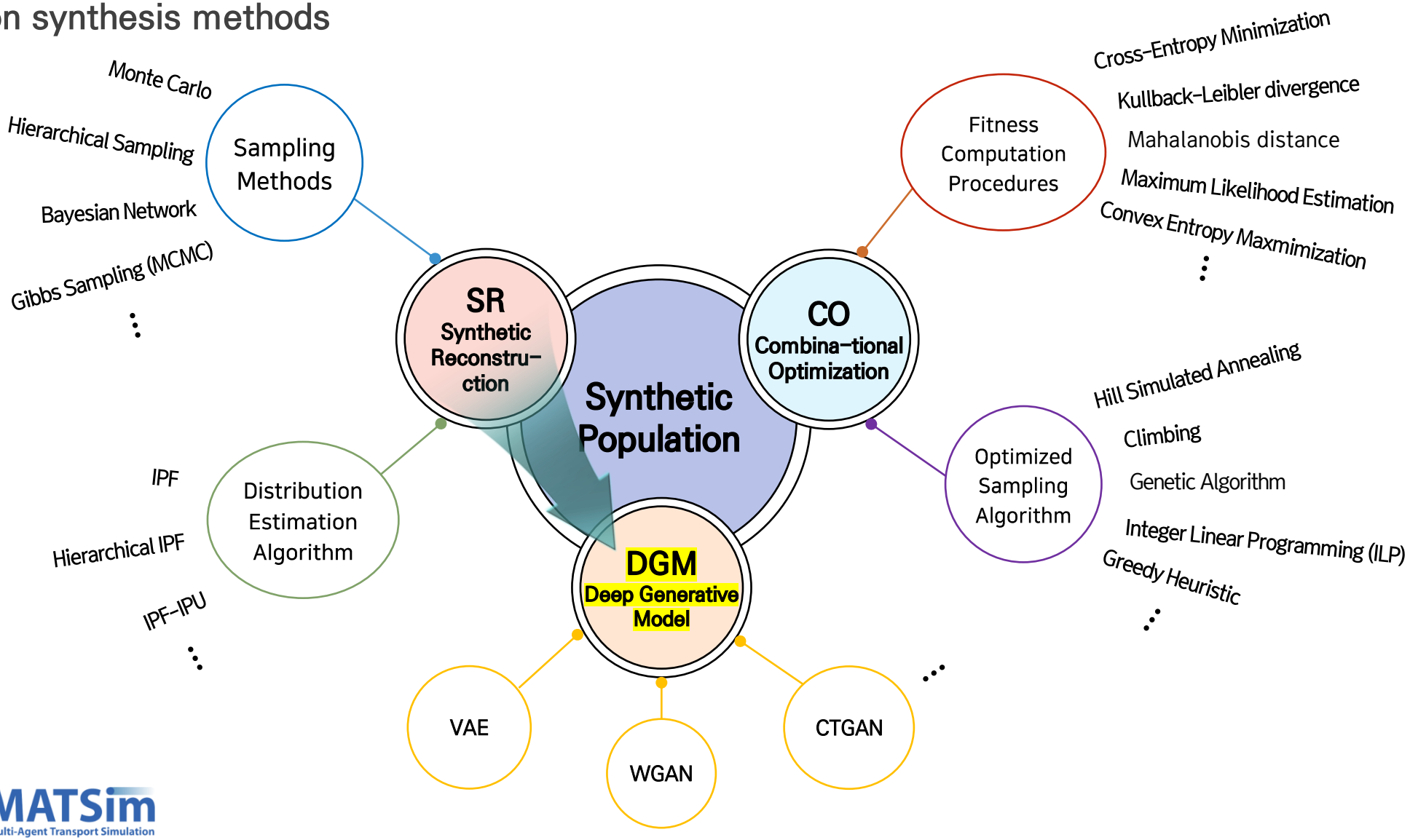
Population synthesis methods: conceptualized procedure

- Since collecting samples from the entire population is not possible, the census offers only aggregated large-scale socio-dem ographic data, while survey data (approximately 1~5%) becomes crucial to capture individual-level attributes.
- The major drawback of the existing approach is its reliance solely on provided sample data, making it incapable of reproducing unseen samples that exist within the population.



Related Works

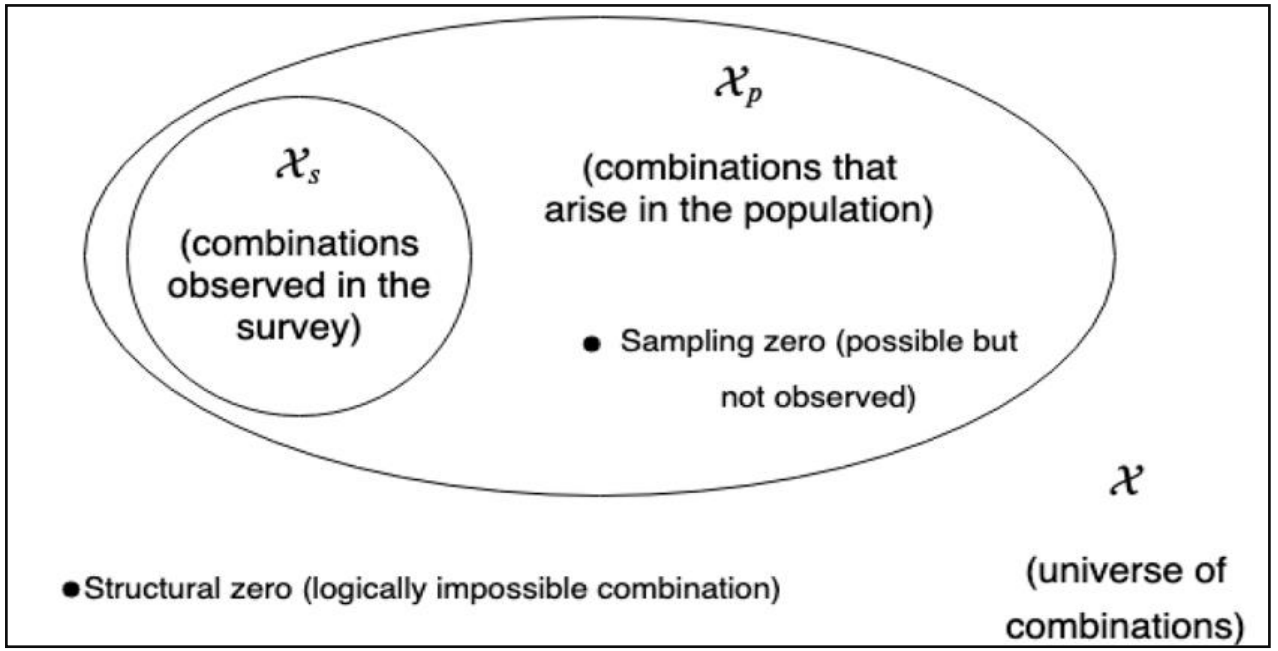
Population synthesis methods



Preliminaries

Structural zeros and sampling zeros in population synthesis

- The DGM model aims to learn the joint probability distribution obtained from survey sample data.
- In the distribution domain, the generated sample types are conceptually distinguished as follows: Structural Zero, which is treated as an infeasible solution, and Sampling Zero, which is unobserved in the sample data but exists within the real population distribution boundary.

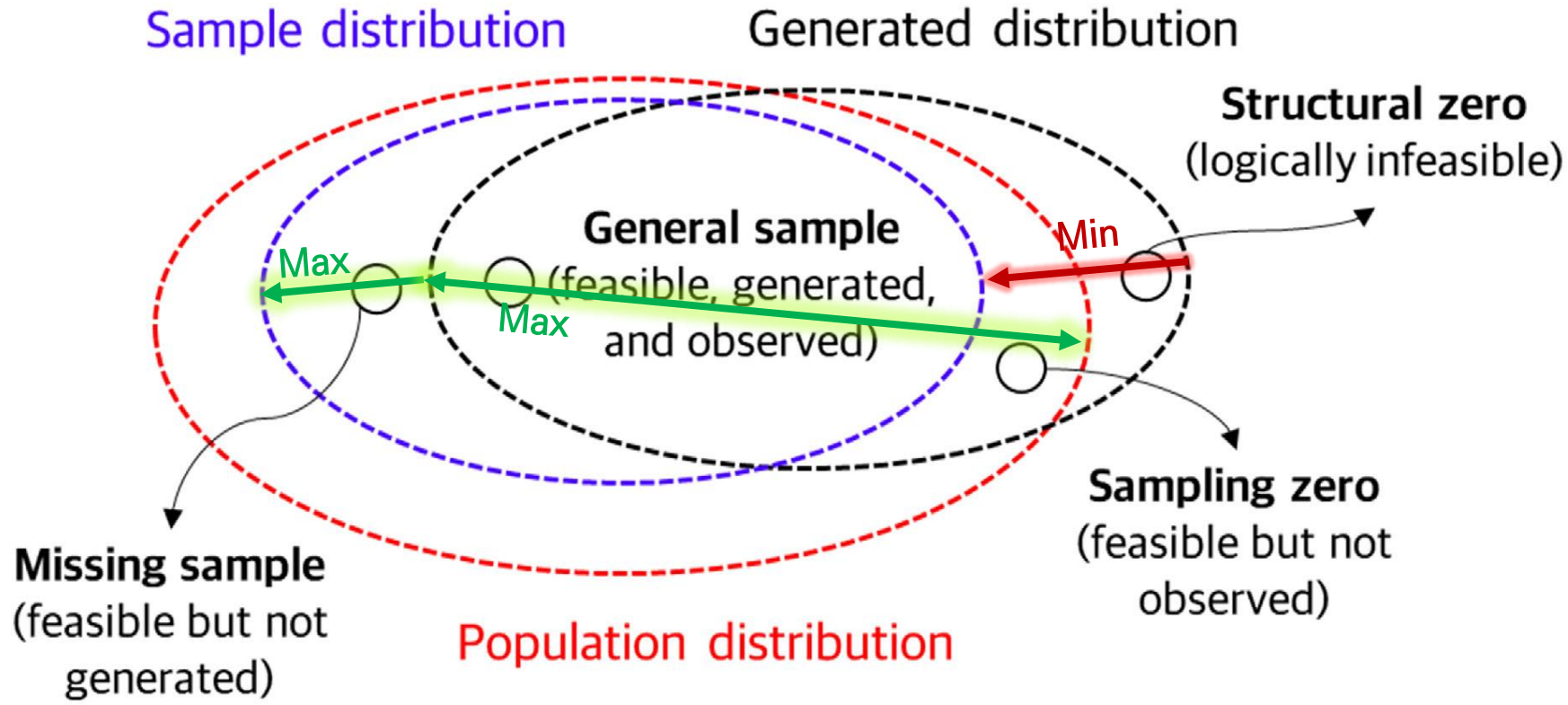


Garrido, Sergio, et al. "Prediction of rare feature combinations in population synthesis: Application of deep generative modelling." Transportation Research Part C: Emerging Technologies 120 (2020): 102787.

Preliminaries

Structural zeros and sampling zeros in population synthesis

- The two distinct issues were tackled by incorporating regularization terms (either Max or Min) into the training process.
- This distinguishing measurement led to a significant improvement in the performance of DGM models for population synthesis.

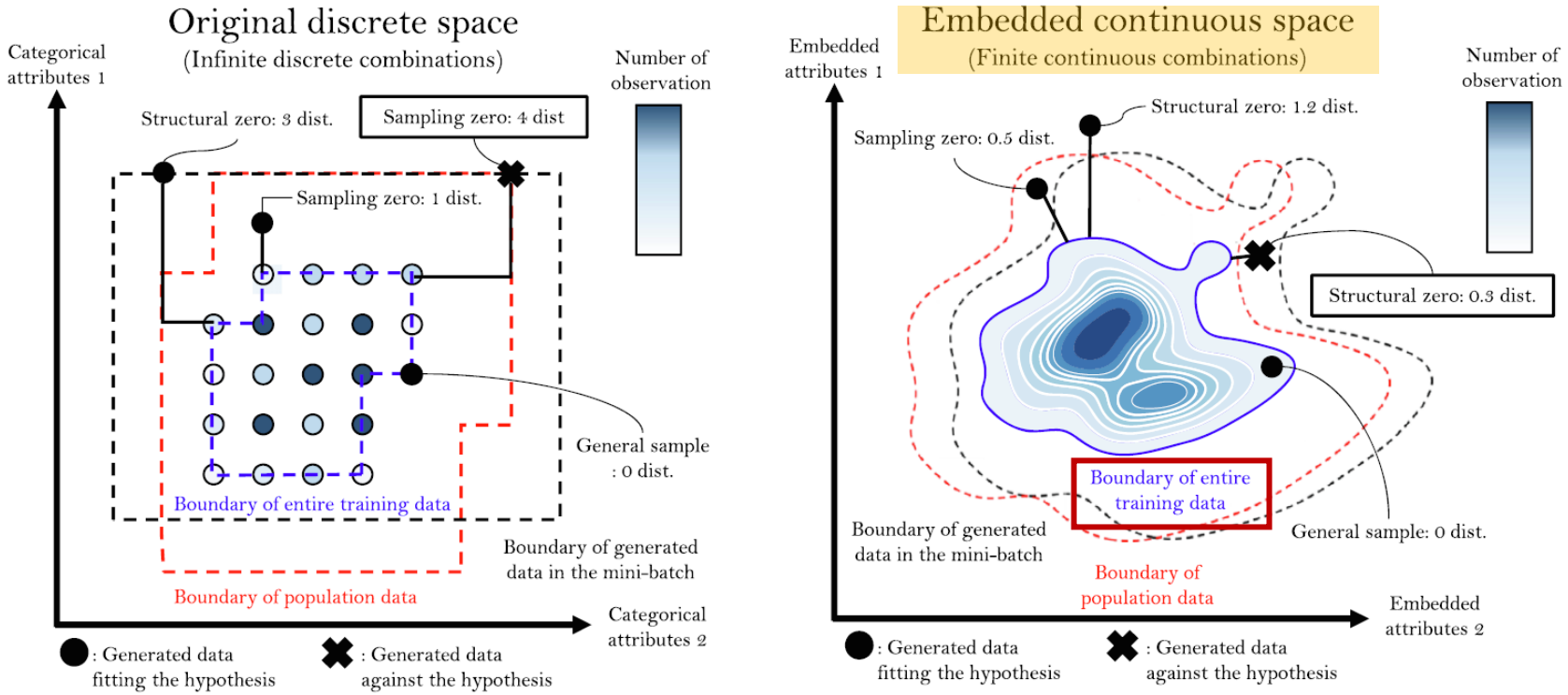


Kim, Eui-Jin, and Prateek Bansal. "A deep generative model for feasible and diverse population synthesis." Transportation Research Part C: Emerging Technologies 148 (2023): 104053.

Preliminaries

Distinguishing structural and sampling zeros

- The novel regularization terms developed by a previous study are designed under the assumption that the sampled zeros have a relatively closer proximity to the training sample distribution boundary, while the structural zeros are observed at greater distances.

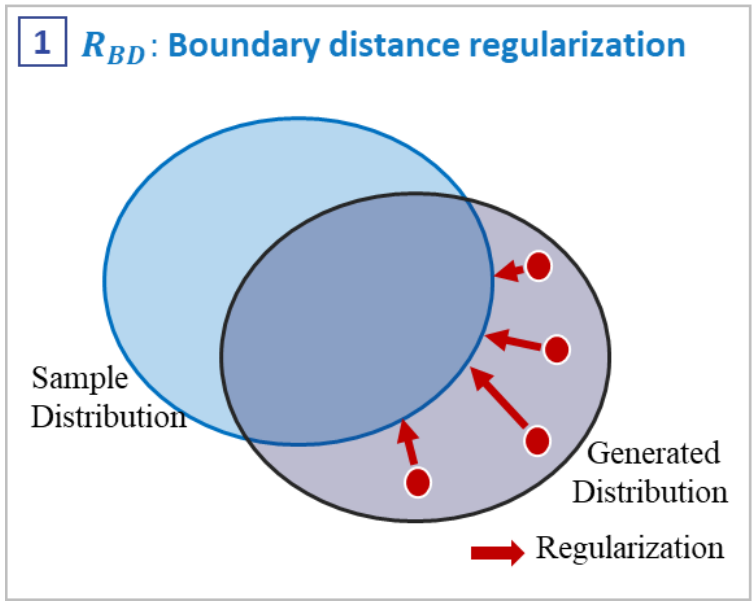


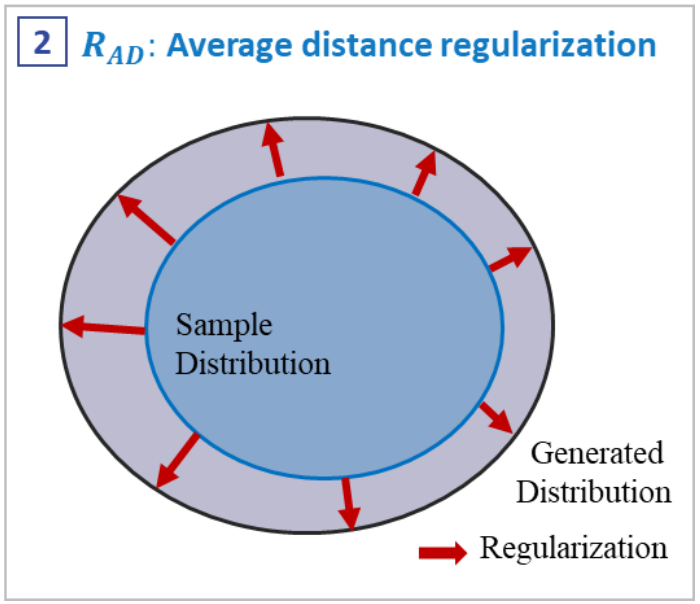
Kim, Eui-Jin, and Prateek Bansal. "A deep generative model for feasible and diverse population synthesis." Transportation Research Part C: Emerging Technologies 148 (2023): 104053.

Preliminaries

Regularization terms (additional loss applied to the generator)

- The $\boldsymbol{R_{BD}}$ suppresses structural zero by minimizing the distance between generated sample and the distribution boundary, whereas the $\boldsymbol{R_{AD}}$ induces sampling zeros by minimizing negative average distribution boundary.
- The proposed terms are added to the generator loss with small loss weight, and these have a trade-off relationship between accuracy and diversity.





$$R_{BD}(\hat{X}, X^S) = \frac{1}{M}\sum_{j=1}^{M} \min_{i\in\{1:N\}, j\in\{1:M\}} (Dist(\hat{X}_j, X_i^S))$$

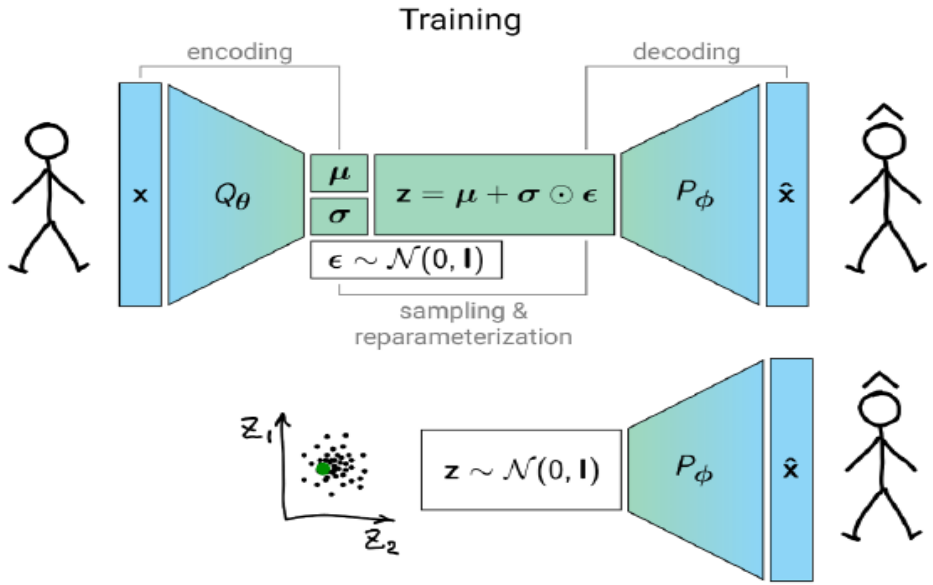
$$R_{AD} = -\frac{1}{NM}\sum_{j=1}^{M}\sum_{i=1}^{N} Dist\,(\hat{X}_j, X_i^S)$$

Preliminaries

Selection of DGM

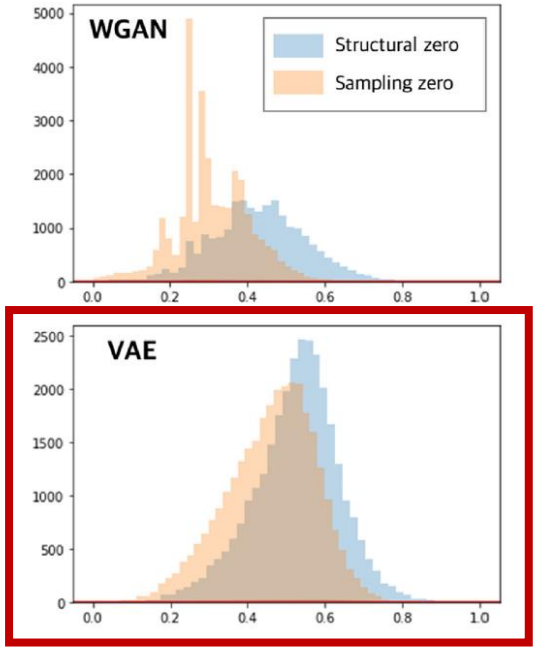
- Performance evaluation of the regularization terms was carried out by measuring the distance of structural and sampling zeros, which serve as indicators of the fidelity of the generated samples.
- There are mainly two generative models applied for population synthesis, which the Variational Autoencoders(VAE) presented in below have a property of smoothening the probability distribution that allows to reproduction of fuzzy, but more diverse samples.

Variational Autoencoders (VAE)



Borysov, S. S., Rich, J., & Pereira, F. C. (2019). How to generate micro-agents? A deep generative modeling approach to population synthesis. Transportation Research Part C: Emerging Technologies, 106, 73-97.

Distance to the boundary of sample distribution

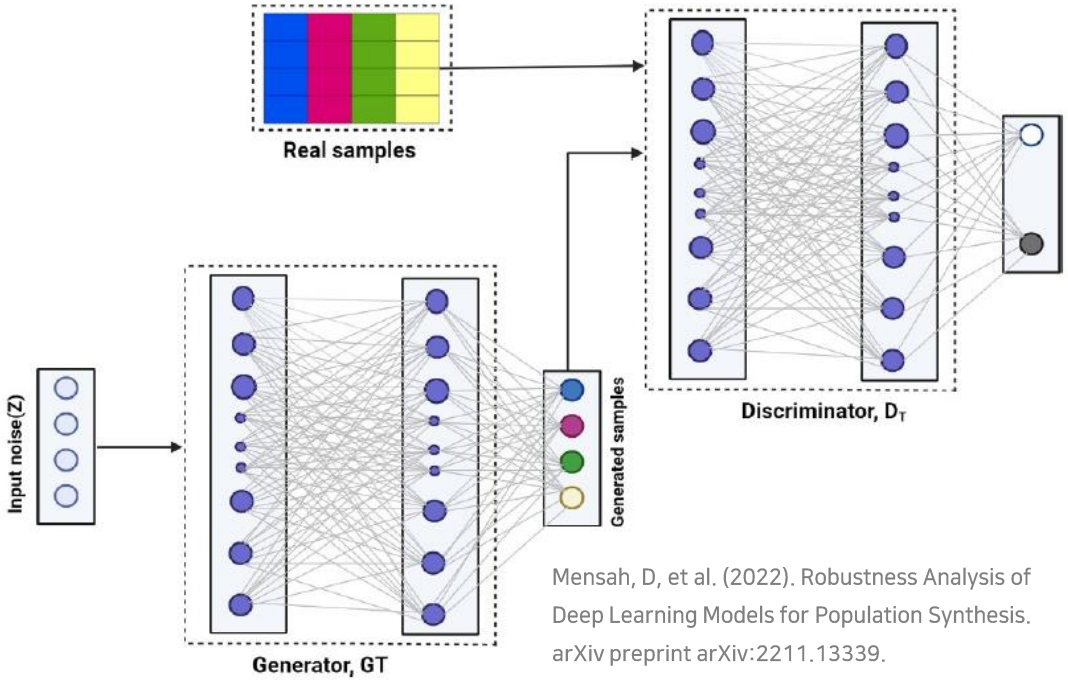


Preliminaries

Selection of DGM

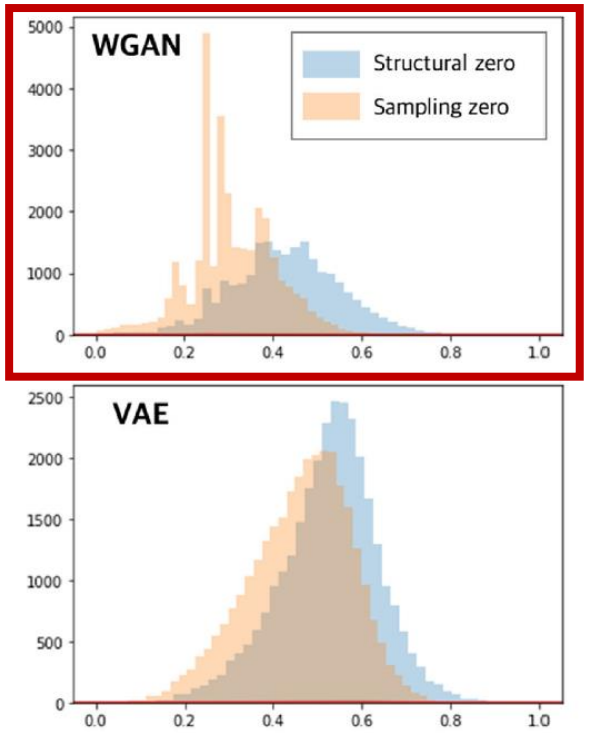
- On the other hand, the Generative Adversarial Networks (GANs) have distinctively closer distances on sampling zeros.
- Owing to their ability to directly match probability distribution, it closely resembles the sample data with high accuracy.
- Additionally, their stability was further enhanced through the implementation of the Wasserstein GAN and gradient penalty.

Generative Adversarial Network (GAN)



Mensah, D, et al. (2022). Robustness Analysis of Deep Learning Models for Population Synthesis. arXiv preprint arXiv:2211.13339.

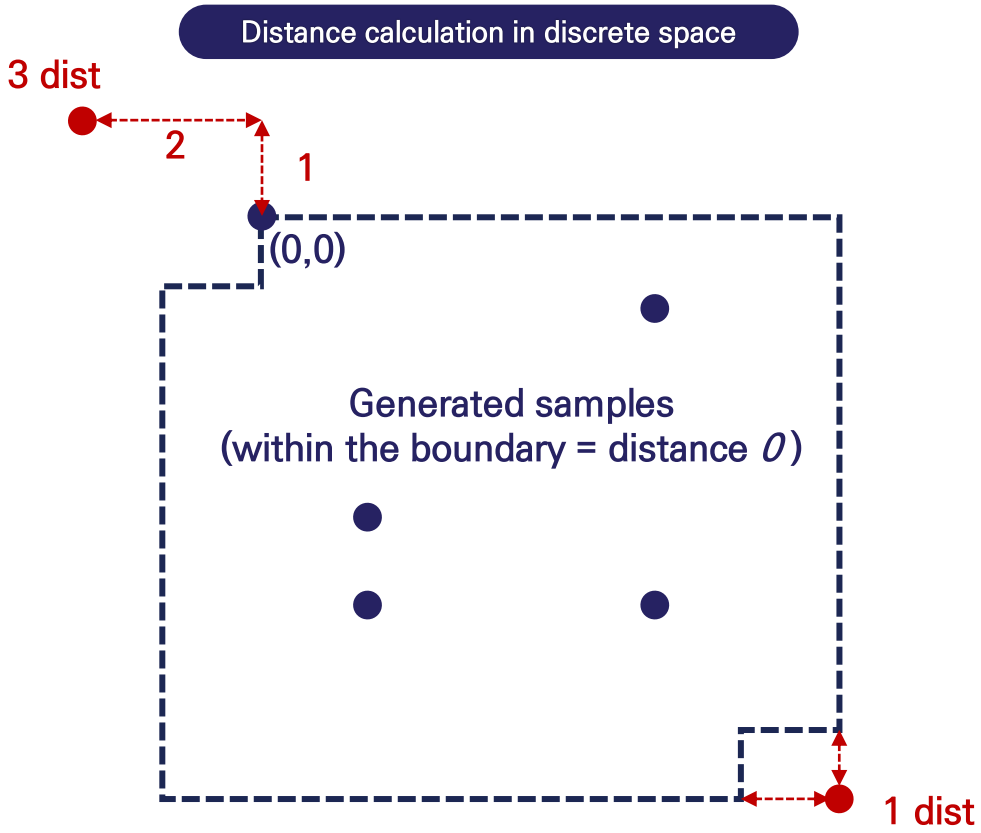
Distance to the boundary of sample distribution

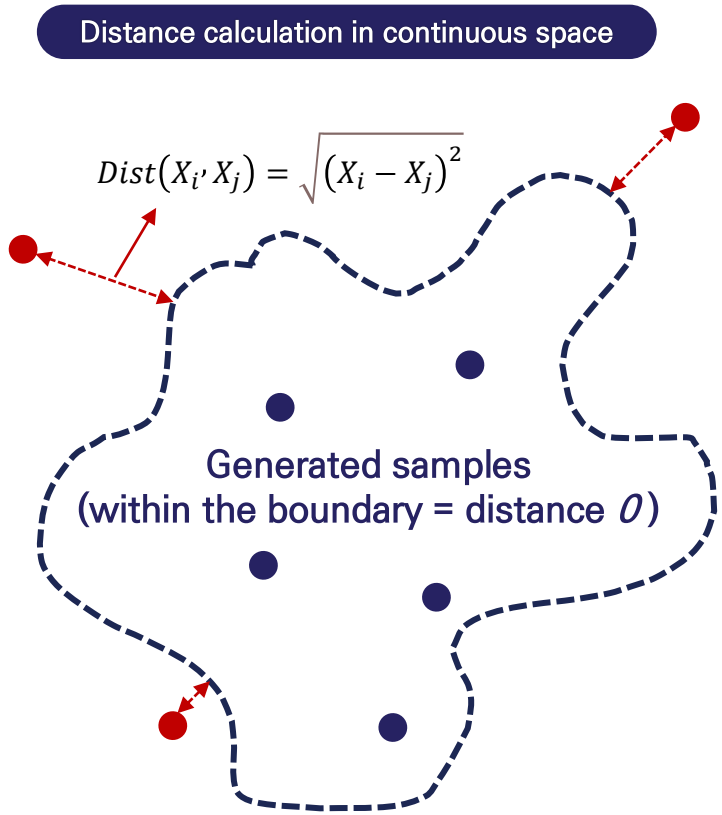


Methodology

Distance measurement

- The calculation of the distance between the generated samples and the distribution boundary of the sample data is straightforward.
- In continuous space, the Euclidean distance is directly employed, while in discrete space, the measurement is based on grid distance.

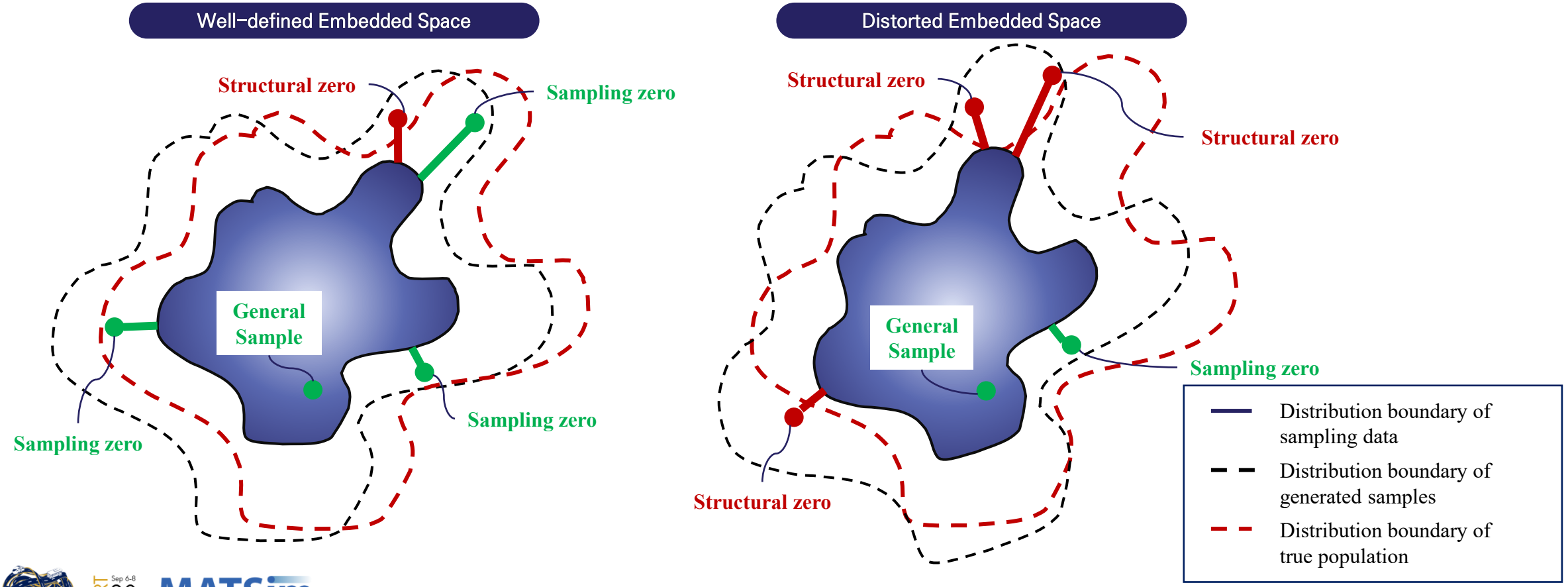




Methodology

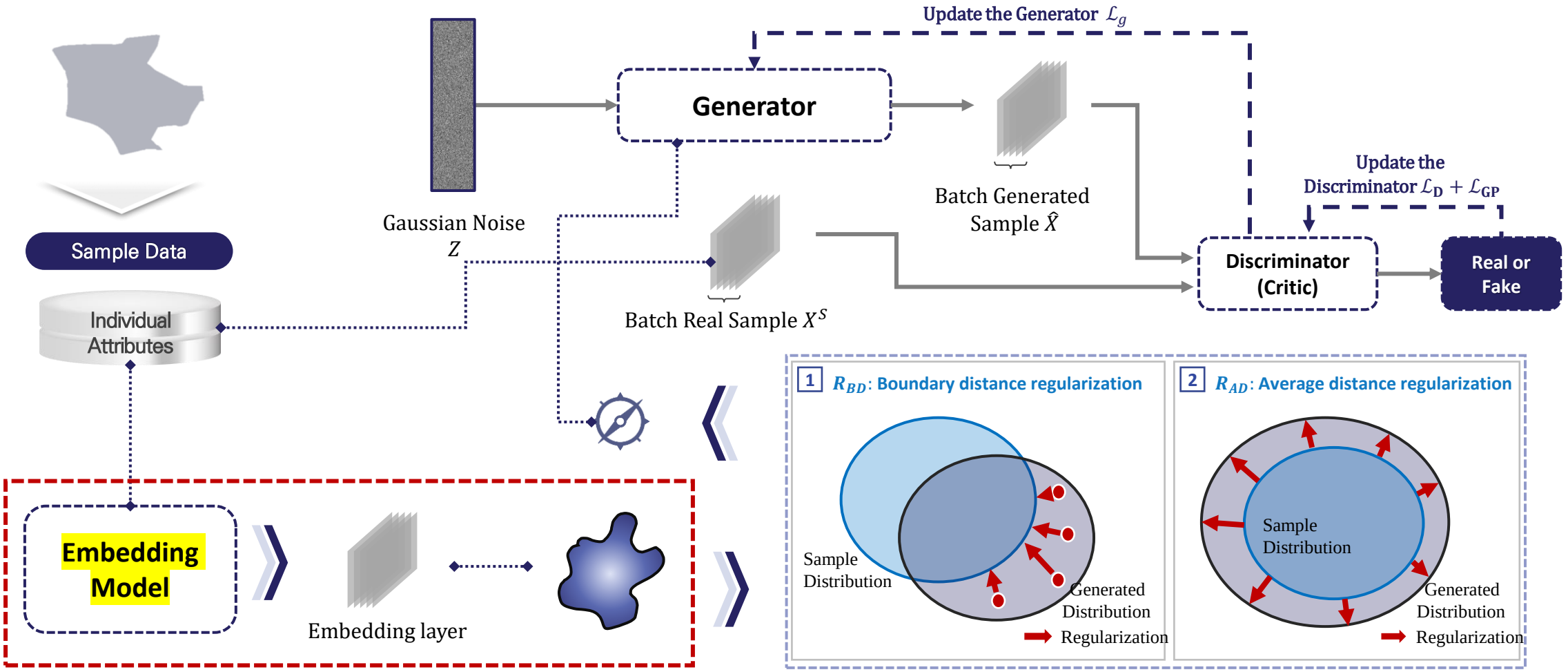
Hypothesis

- Since the embedding space can be formed differently, the tiny distortion from the desired embedding space could lead to a fatal error when distinguishing the structural and sampling zeros.



Methodology

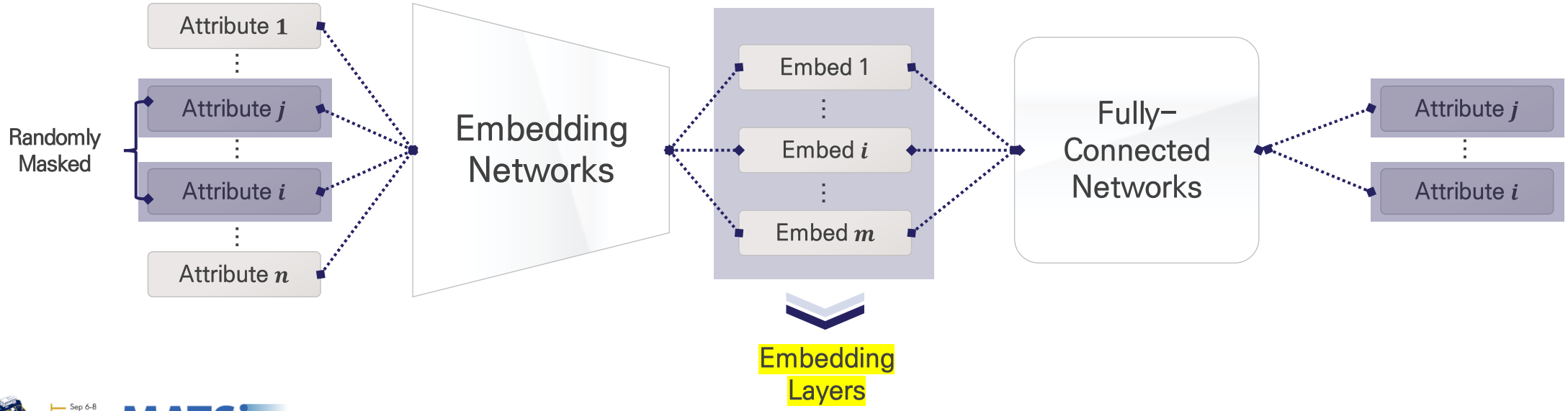
- The overall framework of proposed methodology



Methodology

Embedding space

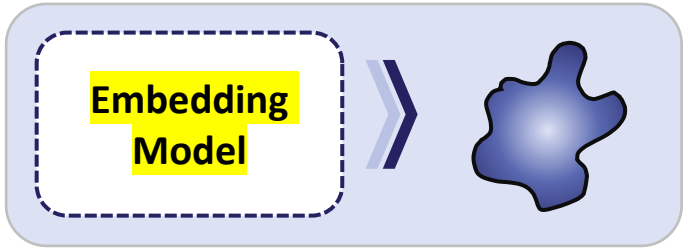
- In order to measure the structural and sampling zeros effectively, the discrete space is converted to embedding space for an accurate distance measurement.
- In this matter, the profound concept of pretraining BERT was adopted, as it learns the contextualized representations of words by predicting the masked words within a sentence. (e.g., age of person 15 with no jobs are likely to commute for school)
- The high dimensional representation of one-hot encoded vectors can be converted into low dimensional space while retaining the information and contextual relationship between the individual attributes.



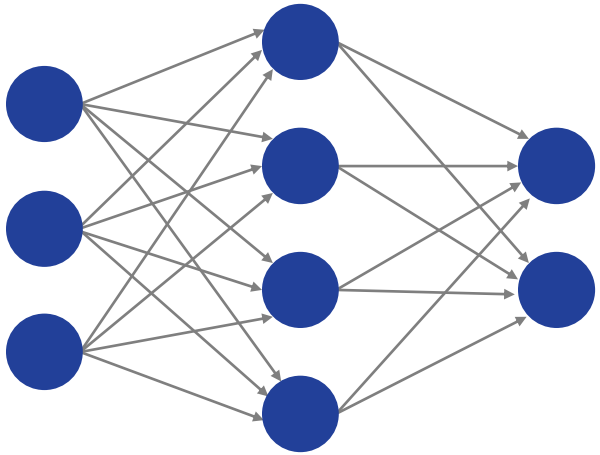
Methodology

Proposed embedding models

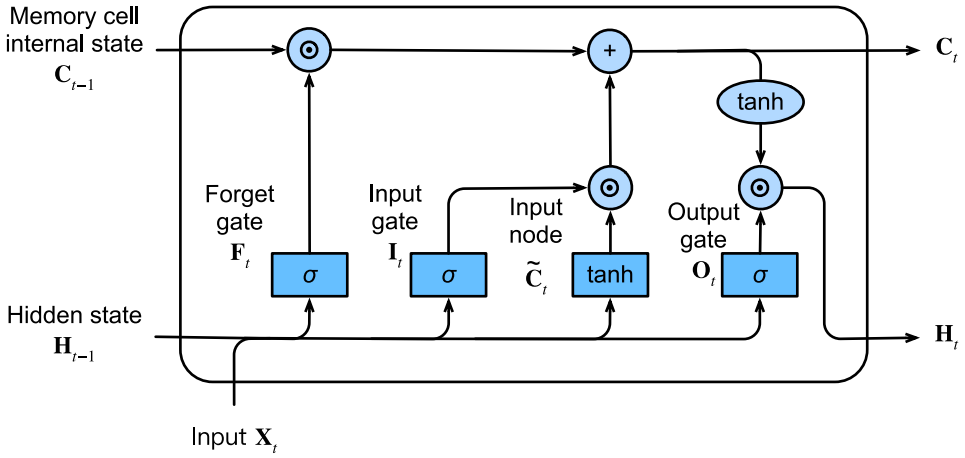
- To enhance the performance of the embedding networks, we leverage well-established NLP models known for effectively capturing contextual relationships and semantic information.
- The obtained embedding space is used for training the WGAN model to test on pre-defined regularization terms



Dense layers



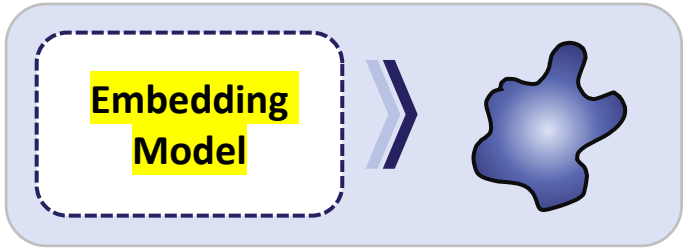
LSTM



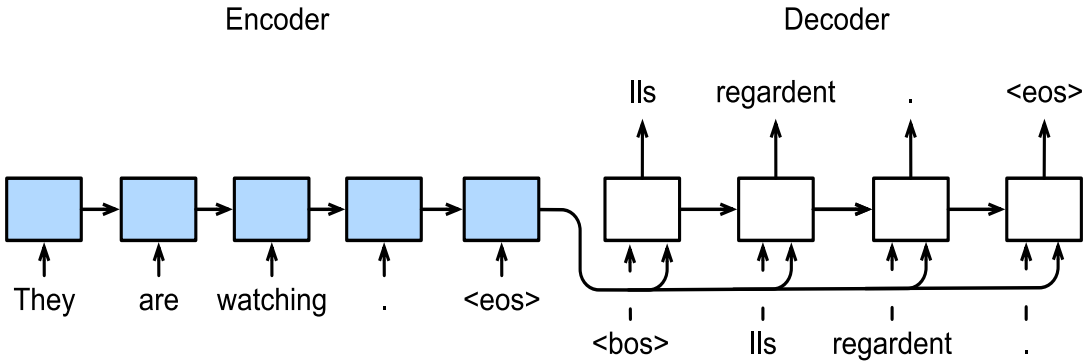
Methodology

Proposed embedding models

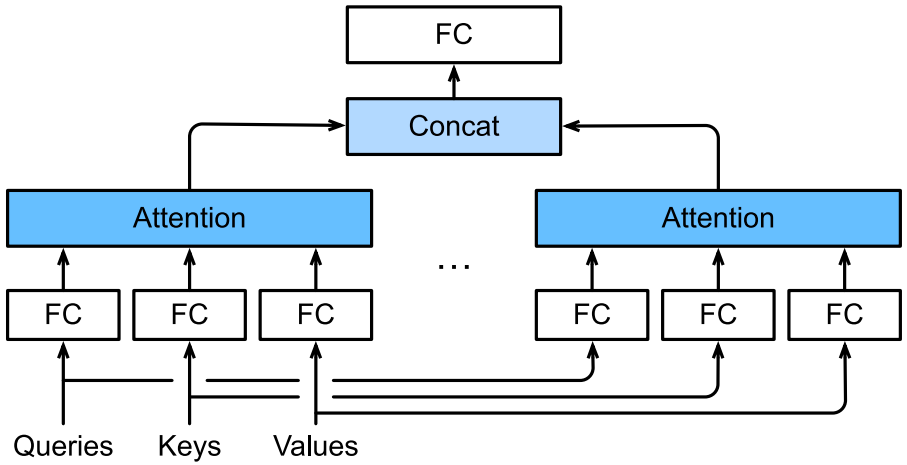
- To enhance the performance of the embedding networks, we leverage well-established NLP models known for effectively capturing contextual relationships and semantic information.
- The obtained embedding space is used for training the WGAN model to test on pre-defined regularization terms



Seq2Seq



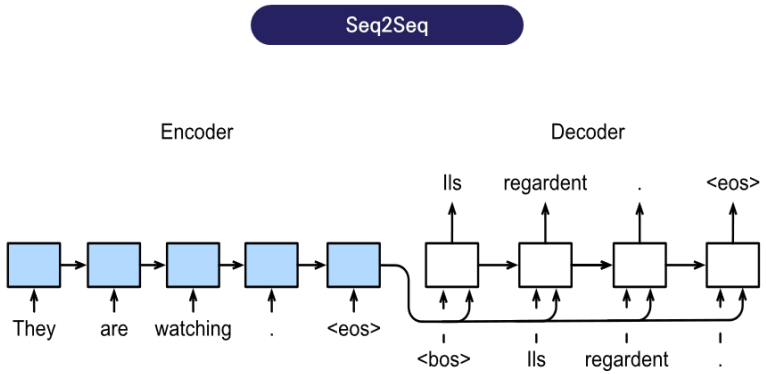
Multi-head Attention

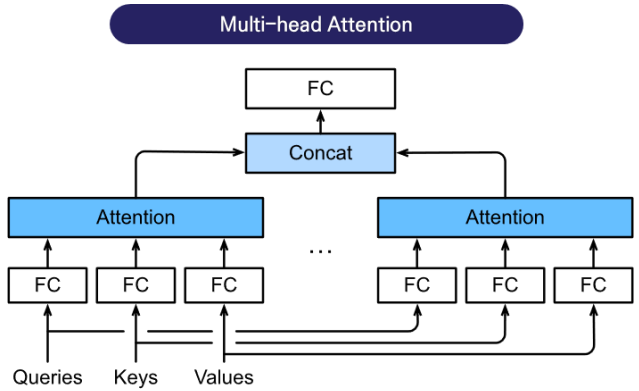


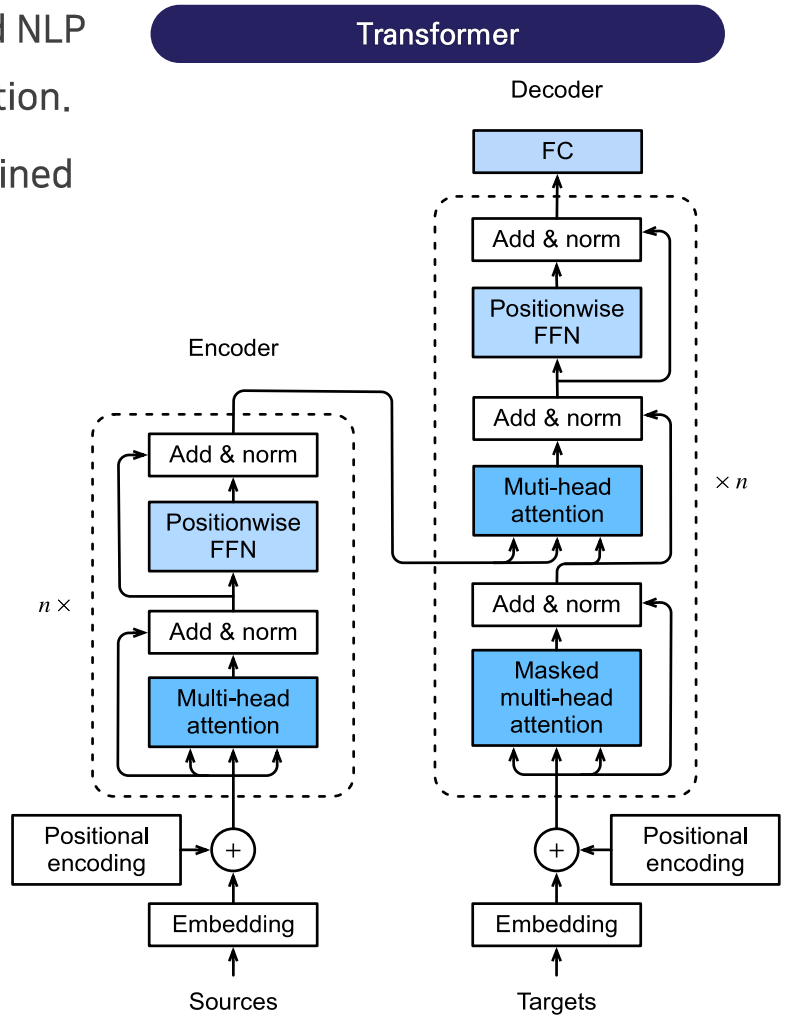
Methodology

Proposed embedding models

- To enhance the performance of the embedding networks, we leverage well-established NLP models known for effectively capturing contextual relationships and semantic information.
- The obtained embedding space is used for training the WGAN model to test on pre-defined regularization terms



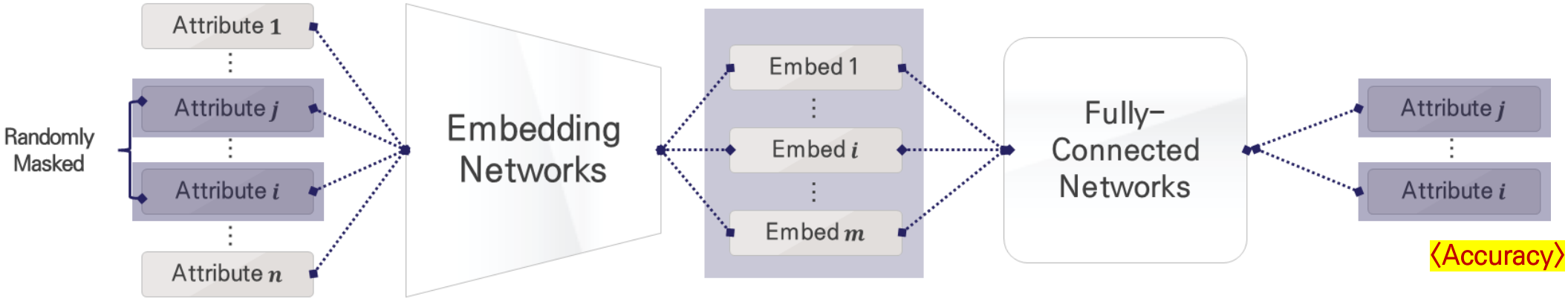




Performance Results

Masked attribute prediction result

- We evaluate the well-defined embedding space based on the masked attribute prediction performance, measured by accuracy.

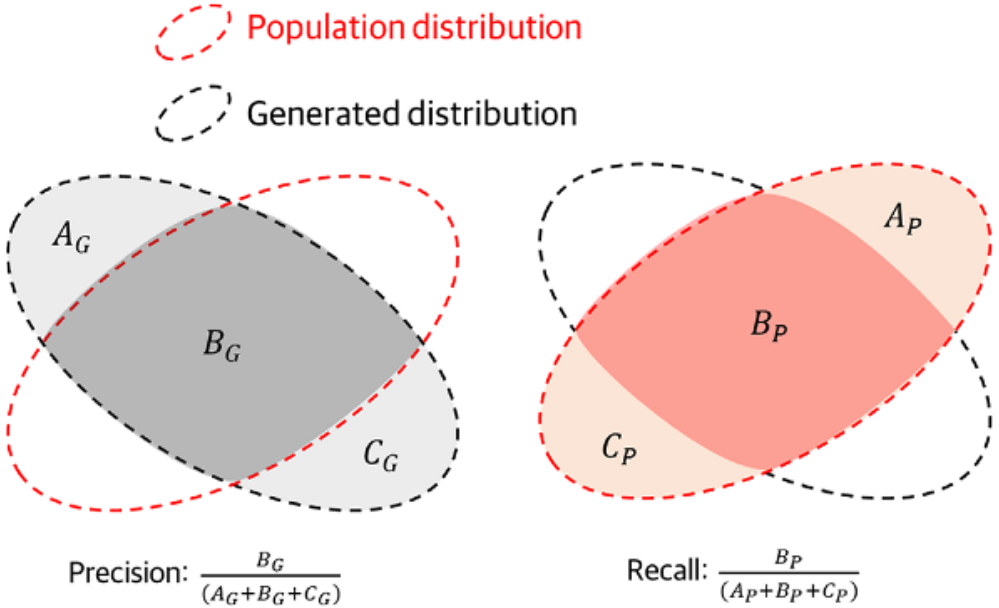


	Dense	LSTM	Seq2Seq	Multi-head Attention	Transformers
Masking 1	0.818	0.844	0.834	0.843	0.846
Masking 2	0.809	0.846	0.834	0.842	0.849
Masking 3	0.806	0.847	0.834	0.843	0.850
Masking 4	0.767	0.849	0.833	0.842	0.849
Masking 5	0.729	0.848	0.832	0.842	0.850
Test Loss	2.563	1.691	1.863	1.778	1.669

Performance Results

Population synthesis via DGM result

- The baseline method is the raw sample data drawn from the total population.
- The WGAN-gradient penalty model is applied for the performance result in terms of accuracy and diversity, summarized by the F1 score.



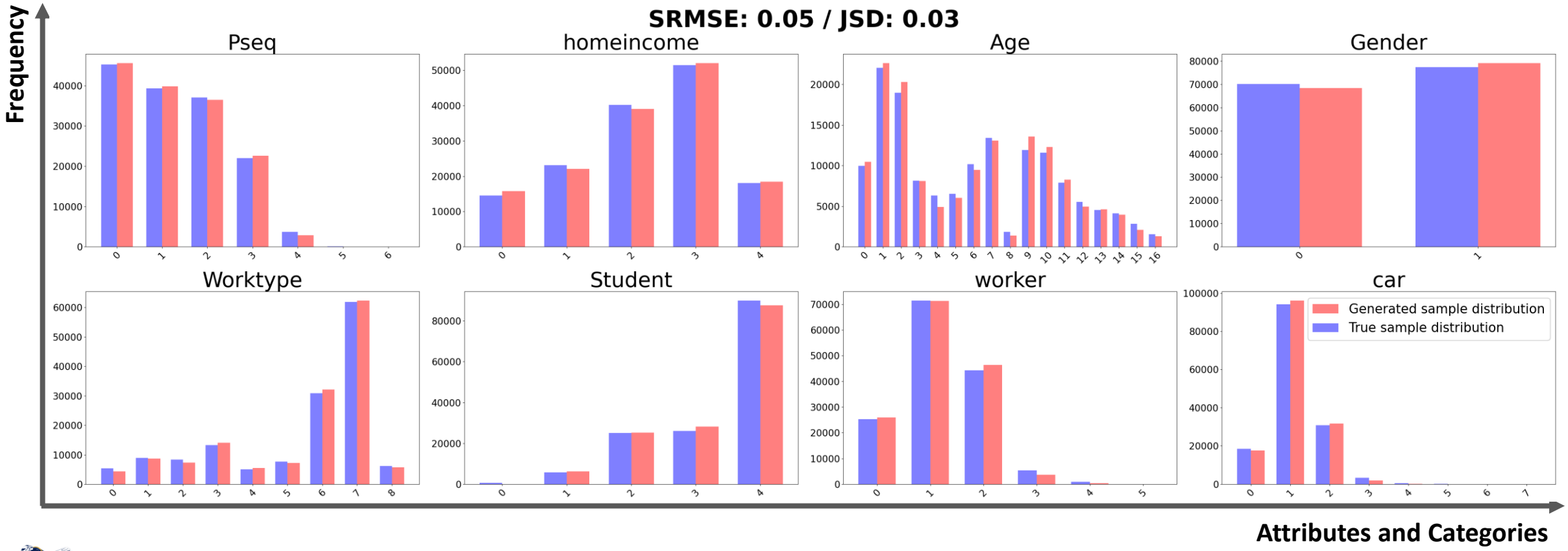
$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

	Precision	Recall	F1 Score
Baseline	1	0.604	0.753
Discrete space (No reg. is applied)	0.765	0.823	0.793
Dense layers	0.860	0.758	0.806
LSTM	0.867	0.749	0.804
Multi-head Attention	0.878	0.773	0.822
Seq2Seq	0.890	0.781	0.832
Transformer	0.898	0.781	0.838

Performance Results

Marginal distribution comparison

- We adopted the well-known methods for distribution similarity measurements, which refers to SRMSE(Standard Root Mean Squared Error) and JSD(Jensen-Shannon Distance) to compare results of real population and generated sample



Conclusion and Future study

Implication of the study

- This study shows the robustness of the regularization terms for reducing structural zeros and inducing sampling zeros for DGM developed by the prior study.
- The primary goal of this study is to create a synthetic population that matches the real population in terms of accuracy and diversity.
- Since utilizing the embedding space is essential to apply the regularization terms effectively, we propose a sophisticated procedure when converting discrete data is crucial to enhance accuracy and diversity in population synthesis.
- Owing to the state-of-the-art models used for the NLP domain, the embedding space extracted by the Transformer model showed the best performance and enhanced the performance by more than 5% compared to the discrete method.
- However, several factors should be discussed for the further improvement and promotion of adoption of DGM in population synthesis:
    - Larger population synthesis considering the unbiased generation of population synthesis
    - Geographical allocation of generated samples

References

- Borysov, S. S., Rich, J., & Pereira, F. C. (2019). How to generate micro-agents? A deep generative modeling approach to population synthesis. Transportation Research Part C: Emerging Technologies, 106, 73-97.
- Castiglione, J., Bradley, M., & Gliebe, J. (2015). Activity-based travel demand models: a primer.
- Choupani, A.-A., & Mamdoohi, A. R. (2016). Population synthesis using iterative proportional fitting (IPF): A review and future research. Transportation Research Procedia, 17, 223-233.
- Galli, E., Cuellar, L., Eidenbenz, S., Ewers, M., Mniszewski, S., & Teuscher, C. (2009). ActivitySim: large-scale agent-based activity generation for infrastructure simulation. Proceedings of the 2009 spring simulation multiconference,
- Garrido, S., Borysov, S. S., Pereira, F. C., & Rich, J. (2020). Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. Transportation Research Part C: Emerging Technologies, 120, 102787.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. Advances in neural information processing systems, 30.
- Kim, E.-J., & Bansal, P. (2023). A deep generative model for feasible and diverse population synthesis. Transportation Research Part C: Emerging Technologies, 148, 104053.
- Luger, B. (2017). Generation of a Synthetic Population for MATSim Models Using Multidimensional Iterative Proportional Fitting and Discrete Choice Models Master's Thesis. Graz: TU Graz.

hEART 2023 Sep 6-8

MATSim
Multi-Agent Transport Simulation

Thank you

dh.kwon@kaist.ac.kr
Donghyun Kwon

inhi.kim@kaist.ac.kr
Inhi Kim

KAIST