

Innovations in Urban Computing: Uncertainty Quantification, Data Fusion, and Generative Urban Design

by

Qing Yi Wang

BASc, University of Toronto (2018)

S.M., Massachusetts Institute of Technology (2020)

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Transportation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2024

© Qing Yi Wang 2024. All rights reserved.

The author hereby grants to MIT a nonexclusive, worldwide, irrevocable, royalty-free license to exercise any and all rights under copyright, including to reproduce, preserve, distribute and publicly display copies of the thesis, or release the thesis under an open-access license.

Authored by: Qing Yi Wang

Department of Civil and Environmental Engineering

January 10 2024

Certified by: Jinhua Zhao

Professor of Cities and Transportation

Thesis Supervisor

Accepted by: Heidi Nepf

Donald and Martha Harleman Professor of Civil and Environmental Engineering

Chair, Graduate Program Committee

Innovations in Urban Computing: Uncertainty Quantification, Data Fusion, and Generative Urban Design

by

Qing Yi Wang

Submitted to the Department of Civil and Environmental Engineering
on January 10 2024, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Transportation

Abstract

Today, urban computing has emerged as an interdisciplinary field connecting data science and urban planning, reflecting the growing integration of urban life with advanced computational methods. Urban computing has particularly benefited from deep learning owing to the spatiotemporal and multi-modal nature of data emerging from urban systems. Deep learning models have not only boosted predictive accuracies beyond traditional models, but also adeptly handled unstructured data. However, the application of deep learning in urban system analysis has a lot of challenges. Within the vast and complex scope of urban computing combined with deep learning, this dissertation zooms in on three emerging issues: uncertainty quantification, data fusion, and generative urban design while focusing on transportation systems and urban planning applications.

The first part of this dissertation proposes a framework of probabilistic graph neural networks (Prob-GNN) to quantify spatiotemporal uncertainty. This Prob-GNN framework is substantiated by deterministic and probabilistic assumptions and empirically applied to predict Chicago’s transit and ridesharing demand. Prob-GNNs can accurately predict ridership uncertainty, even under significant domain shifts such as the COVID-19 pandemic. Among the family of Prob-GNNs, two-parameter distributions (e.g., heteroskedastic Gaussian) achieve the highest predictive performance, which is 20% higher in log-likelihood and 3-5 times lower in calibration errors compared to the one-parameter baseline distributions.

The second part addresses data fusion, in which a theoretical framework of deep hybrid models (DHM) was created to combine the numeric and imagery data for travel behavior analysis. DHM aims to enrich the family of hybrid demand models using deep architectures and enable researchers to conduct associative analysis for sociodemographics, travel decisions, and generated satellite imagery. Empirically, DHM is applied to analyze travel mode choice using the Chicago MyDailyTravel Survey as the numeric inputs and the satellite images as the imagery inputs. DHM can construct latent spaces that significantly outperform classical demand and deep learning models in predicting aggregate and disaggregate travel behavior. Such latent

spaces can also be used to generate new satellite images that do not exist in reality and compute the corresponding travel behavior and economic information, such as substitution patterns and social welfare.

The last part develops a human-machine collaboration framework for generative urban design and then instantiates the framework with a model trained to generate satellite imagery from a land use text description and a constraint image depicting the unaltered major road networks and natural environments. The trained model can generate high-fidelity, realistic satellite images while retaining control over the land use patterns in generated images with natural language descriptions, producing alternate designs with the same inputs, respecting the built and natural environment, and learning and applying local contexts from different cities.

Thesis Supervisor: Jinhua Zhao

Title: Professor of Cities and Transportation

Acknowledgments

To the people without whom this dissertation would not be possible, I express my sincere gratitude for their support and encouragement. I would like to express my deepest gratitude to my advisor, Professor Jinhua Zhao, for his unwavering support and guidance throughout my doctoral studies. Jinhua is such a role model for everyone in the lab. His contagious enthusiasm and dedication have been and will always inspire me. To my committee members, Professor Sandy Pentland, Professor Joan Walker, and Professor Shenhao Wang, thank you for your involvement and generous advice. I am deeply grateful to Sandy and Joan for being involved in this research and providing their insightful feedback. I have been working with Shenhao since my first semester at MIT, while he was still a PhD candidate. To the extent that this dissertation is successful, it is crucial to acknowledge Shenhao's deep involvement from ideation to writing. Shenhao has also been a mentor in life, who was especially helpful during the COVID years. I can always come to him for advice, whether in academics or in life. My gratitude also goes to Professors Haris Koutsopoulos and Nigel Wilson. Although I did not directly work with them recently, they opened the door to research for me during my first two years at MIT. Their attention to detail and rigorous attitude towards research continue to influence me.

None of this would have been possible without the friends I made at MIT. MIT CEE and JTL Mobility Lab are two communities that enriched my life outside of research. In particular, I would like to thank Olivia for being the best friend who always had my back and cured all my insecurities. She is the most supportive friend anyone can have. Special thanks also go to Baichuan, Yunhan, and Xiaotong, with whom I share so much fond memories of countless meals, hikes, board games, and academic discussions. I cannot ask for a better cohort to share my time at MIT with. Then to Yiyun, Xiaoyu, Dingyi, Yuzhu, Yen-Chu, Ruben, John, and many others for creating such a fun and intellectually stimulating environment.

I am also greatly indebted to my family members: my mother Min, and grandparents Langui and Zhuanmo. Being the only child of an only child is not easy, but

we have made it. Regardless of the ups and downs, my family members have always been my biggest supporters. Finally, to my second biggest supporters, Kaiyuan, and his family. PhD life can be very stressful at times, but doing it together made it that much easier. Thank you for always being patient and keeping me sane.

Thank you all for making this journey possible. To conclude, as Jinhua would tell us: “Now go out to the world and make an impact!”. Cheers to a great five years, and to many more great years ahead.

Contents

1	Introduction	15
1.1	Background	15
1.2	Dissertation Overview	17
1.2.1	Uncertainty Quantification	17
1.2.2	Data Fusion	18
1.2.3	Generative Urban Design	19
2	Uncertainty Quantification: Probabilistic graph neural networks for uncertainty quantification of short-term demand prediction	21
2.1	Background	21
2.2	Literature Review	23
2.2.1	Travel Demand Prediction with Deep Learning	23
2.2.2	Uncertainty Quantification Methods	24
2.3	Theoretical Framework	26
2.3.1	Probabilistic Assumptions \mathcal{G}	29
2.3.2	Deterministic Assumptions in \mathcal{F}	31
2.3.3	Specific Examples: Gaussian Distributions and Mean-Variance Estimation	32
2.3.4	Evaluation	33
2.4	Case Study	34
2.4.1	Data	34
2.4.2	Experiment Setup	36
2.5	Results and Discussion	38

2.5.1	Significance of Probabilistic and Deterministic Assumptions	39
2.5.2	Implications of Probabilistic Assumptions	42
2.5.3	Model Performance under System Disruption	44
2.5.4	Spatiotemporal Uncertainty	46
2.6	Conclusion	47
3	Data Fusion: Deep hybrid models to combine numerical data and satellite imagery for travel behavior analysis	51
3.1	Background	51
3.2	Literature review	54
3.2.1	Travel demand modelling	54
3.2.2	Computer vision and urban computing	55
3.3	Theory	57
3.3.1	General framework of deep hybrid models	57
3.3.2	Mixing operator: supervised autoencoders	59
3.3.3	Behavioral predictor	63
3.3.4	Deriving economic information from generated satellite imagery	65
3.4	Experiment design	67
3.4.1	Data	67
3.4.2	Model training	68
3.5	Results	69
3.5.1	Predictive performance	69
3.5.2	Navigating the latent space	72
3.5.3	Deriving economic information for generated satellite imagery	74
3.6	Conclusion	79
4	Generative Urban Design: Human-guided automatic urban design via diffusion models	83
4.1	Background	83
4.2	Literature Review	85
4.2.1	Urban Imagery in Planning and Design	85

4.2.2	Image Generation Models	86
4.2.3	Generative Urban Design	88
4.3	The Human-machine Collaboration Vision	90
4.4	Problem Framing	91
4.5	Methodology	93
4.5.1	Feature Extraction	94
4.5.2	Generative Urban Design	97
4.6	Dataset	98
4.7	Model Evaluation	99
4.8	Model Demonstration	100
4.8.1	Goals	100
4.8.2	Ablation Study	109
4.9	Discussion, Limitations, and Future Work	109
4.10	Conclusion	114
5	Conclusion	117
5.1	Summary	117
5.2	Future Research	120
5.2.1	Uncertainty Quantification	120
5.2.2	Data Fusion: Deep Hybrid Models	121
5.2.3	Generative Urban Design	123
A	Additional Details of Deep Hybrid Models	125
A.1	Specific model design in disaggregate analysis	125
A.2	Effects of mixing and sparsity hyperparameters on prediction and im- age reconstruction	126
A.3	Two-directional imagery regeneration	129
B	Denoising Diffusion Probabilistic Models and Latent Diffusion Mod- els	131
B.1	Denoising Diffusion Probabilistic Models	131

B.2 Latent Diffusion Models (Stable Diffusion)	133
C Generative Design - User Study	135

List of Figures

2-1	Spatial Autocorrelation (Moran’s I) of CTA Rail and Ridesharing Data.	35
2-2	Average Daily Ridership of CTA Rail and Ridesharing (T1: Immediately Before; T2: Stay-at-home; T3: Initial Recovery; T4: Steady Recovery).	37
2-3	Proposed model architecture of Probabilistic Graph Neural Networks.	38
2-4	Training Curves. Top: GCN; Bottom: GAT	39
2-5	Calibration Plot of GCN Models (a) CTA rail and (b) Ridesharing.	41
2-6	Spatiotemporal Uncertainty: Standard deviations of estimated CTA Rail station tap-ins in the 15-minute periods starting at 8 A.M. (morning peak), 1 P.M. (midday), 5 P.M. (afternoon peak), and 9 P.M. (evening).	46
3-1	Diagram of a deep hybrid model (Model 4-6 in Table 3.1)	53
3-2	Supervised Autoencoders	59
3-3	Samples of satellite images used in the experiment.	67
3-4	Two-step training	68
3-5	Non-zero coefficients from sociodemographics and urban imagery in Model 3 with various θ values.	71
3-6	Performance of predicting sociodemographics in Model 6 with various λ values	73
3-7	Image samples of cluster centers	73
3-8	Spatial distribution of clusters	74
3-9	tSNE visualization of latent space	75

3-10	Satellite images of three representative census tracts	76
3-11	One-directional image generation	76
3-12	Economic information of images generated along one direction (Southern to Central Chicago: $\tilde{z} = z_s + a_1(z_c - z_s)$)	77
3-13	Two-directional image generation	78
3-14	Two generated satellite images	79
3-15	Economic information of images generated along two directions	79
4-1	The human-machine collaboration planning process.	91
4-2	Two-stage model training	93
4-3	Sample Training Pairs	94
4-4	Symbols in the Design Constraint Image	95
4-5	Dataset Coverage	98
A-1	Quality of image reconstruction with various λ values	128
A-2	Another example of two-directional image regeneration	129
C-1	Generative Urban Design - User Study Part 1	136
C-2	Generative Urban Design - User Study Part 2	137

List of Tables

2.1	Model Design of the Probabilistic Graph Neural Networks Framework	28
2.2	Three Categories of Evaluation Metrics	33
2.3	Prob-GNN Model Performance (Test Period: Immediately Before - March 2 to March 15, 2020)	40
2.4	Model Generalization under System Disruption	44
3.1	Design of the mixing operator in deep hybrid models	58
3.2	Predictive performance	70
3.3	Sociodemographics of three representative census tracts	75
4.1	User study scores of generated and real satellite images (Min score:1, Max score:5)	99
4.2	True vs. generated satellite imagery	101
4.3	Generations over different land use combinations	103
4.4	Balance between creativity and design descriptions	104
4.5	Adherence to different types of design constraints	106
4.6	Sequentially adding design constraints to generations	107
4.7	Local urban textures in different cities	108
4.8	Ablation study of design descriptions	110
5.1	Limitations and Future Work of Generative Urban Design	123
A.1	Effects of the mixing hyperparameter λ in Model 4	127
A.2	Effects of the sparsity hyperparameter θ in Model 4	128

Chapter 1

Introduction

1.1 Background

Urban analytics involves collecting, analyzing, and interpreting large amounts of data related to urban areas to improve understanding, decision-making, and management of cities. It uses analytical techniques and tools, including statistical analysis, machine learning, and geographic information systems (GIS), to study diverse aspects of urban life, such as transportation, housing, social interactions, and sustainable living. Urban analytics aims to derive insights that can enhance the quality of life, and the urban environment's sustainability and resilience.

Deep learning, a revolutionary technology in artificial intelligence, has transformed how we perform analytics. Deep learning is a subset of machine learning, distinguished by its ability to learn from vast amounts of data through neural networks that mimic the structure and functions of the human brain. These networks consist of multiple layers of interconnected nodes or neurons, each capable of recognizing increasingly complex patterns and features in the data. The inception of deep learning dates back to the mid-20th century with the development of the perceptron, but it was not until the last couple of decades that it gained significant momentum, thanks to advancements in computational power and data availability. A few pivotal moments in the development of deep learning include the introduction of backpropagation, which enabled the training of large networks, convolutional neural networks (CNN),

which are critical in image-related tasks [87], recurrent neural networks (RNN), which are designed for time series and natural language processing [104, 60], generative adversarial training which significantly improved the generation quality of images [45], transformer architecture and attention mechanism which made a real breakthrough in natural language processing and had then become a standard component for large language and vision models [144], and diffusion models which created a new paradigm for generative models [30].

Today, with its predictive power and capability to process unstructured data, deep learning has brought a lot of opportunities to urban analytics. As a result, Urban computing has emerged as an interdisciplinary field connecting data science and urban planning, reflecting the growing integration of urban life with advanced computational methods. This field has benefited immensely from deep learning since the data generated from urban systems are inherently spatiotemporal and multi-modal. Traditionally, gaining insights from spatiotemporal data has relied on statistical models like time series and spatial regression. However, the advent of deep learning has dramatically enhanced the capacity to model and interpret spatiotemporal data with greater accuracy and efficiency. Convolutional neural networks (CNNs) enabled the analysis of grid configurations, while graph neural networks have broadened the scope of spatial representations, making them more general and adaptable. Similarly, for time series data, implementing recurrent neural networks (RNNs) and long short-term memory (LSTM) networks significantly improved the understanding of longitudinal trends and patterns. Nonetheless, it's crucial to remember that spatiotemporal data are abstractions of life. For humans, more intuitive and prevalent forms of representation include images and natural language, which were mostly indecipherable before the era of deep learning. Now, with advanced deep learning techniques, we are not only able to predict sociodemographics using images and natural language but can also generate new, meaningful data. This evolution marks a significant leap in understanding and utilizing urban data, opening up new possibilities for understanding and shaping the urban experience.

Despite the success of deep learning in analyzing urban systems, undoubtedly, a

lot of challenges remain. Given the impossible scope of urban computing with deep learning, this dissertation zooms in on three emerging issues, focusing on transportation systems and urban planning applications.

1.2 Dissertation Overview

Rapid developments in deep learning technologies have led to a surge in urban computing research over the past decade. Yet, due to the complexity of cities, many challenges remain, and this field continues to be in a state of active evolution. This dissertation highlights three areas for improvement in the current urban computing literature: uncertainty quantification, data fusion, and generative urban design. It also suggests a preliminary framework to address each issue. Given the vast scope of these issues, this thesis does not claim to have solved these challenges. Instead, it aims to draw attention to these critical topics and offers an initial approach, hoping to encourage further research and alternative solution strategies. The subsequent sections summarize the research gaps and the contributions made by this dissertation.

1.2.1 Uncertainty Quantification

Recent studies have largely ignored the uncertainty that inevitably exists in any prediction problems while focusing on improving prediction accuracy using deep learning. However, overlooking uncertainty leads to theoretical and practical issues. In theory, travel demand is inherently a random quantity, characterized by a rich family of probability distributions, sometimes even “fat-tail” ones that render a simple average approximation highly inappropriate [174]. In practice, while predicting average demand provides a valuable basis for system design, the lack of uncertainty analysis precludes using deep learning to provide robust real-time service or resilient long-term planning [48, 49]. Since the recent work has demonstrated the outstanding capability of deep learning in modeling spatiotemporal dynamics, it seems timely and imperative to enrich the deterministic deep learning models to capture the spatiotemporal uncertainty of travel demand.

To fill this gap, in Chapter 2, we propose a probabilistic graph neural networks (Prob-GNN) framework to quantify the spatiotemporal uncertainty. This Prob-GNN framework is substantiated by deterministic and probabilistic assumptions and empirically applied to predicting the demand for transit and ridesharing in Chicago. We found that the probabilistic assumptions (e.g. distribution tail, support) have a more significant impact on uncertainty prediction than the deterministic ones (e.g. deep modules, depth). Among the family of Prob-GNNs, the GNNs with truncated Gaussian and Laplace distributions achieve the highest performance in transit and ridesharing data. Even under significant domain shifts, Prob-GNNs can predict the ridership uncertainty stably when the models are trained on pre-COVID data and tested across multiple periods during and after the COVID-19 pandemic. Prob-GNNs also reveal the spatiotemporal pattern of uncertainty, which is concentrated on the afternoon peak hours and the areas with large travel volumes. Our findings highlight the importance of incorporating randomness into deep learning for spatiotemporal ridership prediction. Future research should continue investigating versatile probabilistic assumptions to capture behavioral randomness and further develop methods to quantify uncertainty to build resilient cities.

1.2.2 Data Fusion

In Chapter 3, we create a theoretical framework of **deep hybrid models** to integrate the numeric and imagery data for travel behavior analysis. Classical demand modeling analyzes travel behavior using only low-dimensional numeric data (i.e. sociodemographics and travel attributes) but not high-dimensional urban imagery. In recent years, a lot of studies have used urban imagery sociodemographic and travel behavior analysis. However, the two types of information should be complementary, thus necessitating a synergetic framework to combine them.

Empirically, this framework is applied to analyze travel mode choice using the Chicago MyDailyTravel Survey as the numeric inputs and the satellite images as the imagery inputs. We found that deep hybrid models significantly outperform both classical demand models and deep learning models in predicting aggregate and

disaggregate travel behavior. The deep hybrid models can reveal spatial clusters with meaningful sociodemographic associations in the latent space. The models can also generate new satellite images that do not exist in reality and compute the corresponding economic information, such as substitution patterns and social welfare. Overall, the deep hybrid models demonstrate the complementarity between the low-dimensional numeric and high-dimensional imagery data and between the traditional demand modeling and recent deep learning. They enrich the family of hybrid demand models using deep architecture to create the latent space and enable researchers to conduct associative analysis for sociodemographics, travel decisions, and generated satellite imagery. Future research could address the limitations in interpretability, robustness, and transferability and propose new methods to enrich the deep hybrid models further.

1.2.3 Generative Urban Design

Due to the labor-intensive and time-consuming nature of urban planning, tools that expedite design, communication, and feedback are highly desired. Most models are predictive and descriptive, yielding insights but rarely directly aiding decision-making. With the advent of generative models, especially large language and vision models, communication between humans and machines has become much more accessible. In Chapter 4, we envision a human-machine collaboration framework for urban planning based on recent advances in generative AI models.

To instantiate this framework, we trained a state-of-the-art diffusion model to generate realistic satellite images with a given text description and constraint image depicting the unaltered major road networks and natural environments. One challenge in training large models is always data availability. Large generative models consume a lot of data. However, labeled data in the urban setting is expensive and scarce. Therefore, we use the globally available OpenStreetMap data to promote scalability, applicability, and standardization. The trained model can generate high-fidelity, realistic satellite images while retaining control over generated images with natural language descriptions, producing alternate designs with the same inputs, re-

specting existing infrastructure and natural environments, and learning and applying local contexts from different cities. As this technology is still nascent, we discuss the feedback from practitioners and scholars in the field and point out future research directions to realize the human-machine collaboration vision for urban planning.

Chapter 2

Uncertainty Quantification: Probabilistic graph neural networks for uncertainty quantification of short-term demand prediction

2.1 Background

Uncertainty prevails in urban mobility systems. An enormous number of uncertainty sources range from long-term natural disasters and climate change to real-time system breakdown and to the inherent randomness in travel demand. These uncertainty sources exhibit spatial and temporal patterns: travel demand could drastically change during the holidays or become spatially concentrated in special events (e.g. football games). Traditionally, the spatial uncertainty is analyzed by spatial econometrics or discrete choice models, while the temporal one is by time series models. Deep learning models have recently been designed to capture spatiotemporal correlations, with Long Short Term Memory (LSTM) networks for temporal correlations and convolutional or graph neural networks (CNN, GNN) for spatial dependencies. Deep learning has been shown to significantly improve travel demand prediction accuracy. However,

most studies focus on designing *deterministic* deep learning models to predict the *average* travel demand but largely ignore the uncertainty inevitably associated with any prediction.

Overlooking uncertainty leads to theoretical and practical problems. In theory, travel demand is inherently a random quantity characterized by a rich family of probability distributions, sometimes even “fat-tail” ones that render a simple average approximation highly inappropriate [174]. In practice, while predicting average demand provides a valuable basis for system design, the lack of uncertainty analysis precludes using deep learning to provide robust real-time service or resilient long-term planning [48, 49]. Since recent work has demonstrated the outstanding capability of deep learning in modeling spatiotemporal dynamics, it seems timely and imperative to enrich the deterministic deep learning models to capture the spatiotemporal uncertainty of travel demand.

To address this research gap, we propose a probabilistic graph neural networks (Prob-GNN) framework to quantify the spatiotemporal uncertainty in travel demand. This framework is substantiated by deterministic and probabilistic assumptions: the former refers to the various architectural designs in deterministic neural networks, while the latter refers to the probabilistic distributions with various uncertainty characteristics. The framework is used to compare the two types of assumptions by both uncertainty and point prediction metrics. This study makes the following contributions:

1. We identify the research gap in quantifying uncertainty for travel demand using deep learning. Despite many deep learning techniques for predicting average travel demand, only a limited number of studies have sought to quantify spatiotemporal uncertainty.
2. To fill the research gap, the paper introduces a general framework of Prob-GNN, which comprises deterministic and probabilistic assumptions. This framework is further substantiated by twelve architectures, all of which can quantify the spatiotemporal uncertainty in travel demand.

3. We find that probabilistic assumptions significantly influence uncertainty predictions while only slightly influencing point predictions. Conversely, the deterministic assumptions, such as graph convolution, network depth, and dropout rate, lead to a similar performance in both point and uncertainty predictions when probabilistic assumptions are fixed.
4. We further demonstrate the generalizability of the Prob-GNNs by comparing the performance across multiple shifting domains caused by COVID-19. Although the mean prediction encounters significant prediction errors, the interval prediction is accurate across the domains.

2.2 Literature Review

The literature on spatiotemporal travel demand prediction using deep learning has grown significantly in recent years. Among them, very few studies have sought to understand the predictive uncertainty. This literature review will first examine the latest developments in demand predictions using deep learning and, subsequently, delve into uncertainty quantification.

2.2.1 Travel Demand Prediction with Deep Learning

Researchers have improved spatiotemporal travel demand prediction accuracy using advanced neural network architectural designs [175, 169, 91, 193]. Temporally, LSTM networks and Gated Recurrent Units (GRU) layers are used to analyze the seasonal, weekly, and time-of-day trends [73, 180, 92]. Spatially, the analysis unit is often a station/stop, urban grid, individuals, or census tract. The urban grid is often analyzed with convolutional neural networks (CNN) [175]. Individual travel demand is analyzed by neural networks with behavioral insights for innovative architectural design [152, 156, 158]. More complex urban structures, such as transit lines that connect upstream and downstream stops, can be represented by graphs. Graph convolution networks (GCN) are often used to model such spatial propagation of traffic/demand

flow into adjacent nodes [171, 79, 179, 94]. The baseline GCNs have been expanded by defining multiple graphs to describe the spatial similarity in points of interest (POIs) and transportation connectivity [43, 135]. An alternative to GCN is Graph Attention Network (GAT), which automatically learns the attentional weightings of each node, thus detecting the long-range dependencies without handcrafted weighting mechanisms through adjacency matrices [93, 21].

Despite the abundance of deep learning models, few studies sought to quantify the travel demand uncertainty, typically caused by narrowly defined prediction objectives and methods. The prediction objective - the average ridership - can hardly represent the randomness in travel demand, particularly for distributions with fat tails, in which the outliers deviating from the average value could predominate [174]. The prediction method - the family of deterministic neural networks - is only a specific example of probabilistic neural networks with homogeneous assumptions on the variance. If unaccounted for, demand uncertainty can negatively influence downstream planning tasks [49, 160]. Uncertainty can also propagate and be significantly enhanced at each stage in multi-stage models [189]. Because of its importance, uncertainty has started to attract attention in the field of travel demand predictions using deep learning. Existing studies analyzed heteroskedastic noises in the Gaussian distribution [120, 151], used Monte-Carlo dropouts to approximate model uncertainty [89, 99]. Additionally, the Gaussian assumption is being challenged. Studies modeled irregular distributions with quantile regressions [126], and modeled sparse demand with zero-inflated distributions [194]. Compared to the research in deterministic architectural designs, the current literature on predictive uncertainty in the travel demand realm using deep learning is still relatively thin. A comprehensive framework encompassing the deterministic and probabilistic assumptions is lacking to allow for a thorough comparison.

2.2.2 Uncertainty Quantification Methods

Research on uncertainty quantification has been developed in other deep-learning fields. There are two major categories of uncertainty: data uncertainty and model uncertainty. Data uncertainty refers to the irreducible uncertainty inherent in the data

generation process, while model uncertainty captures the uncertainties in the model parameters. For example, in linear regression, data uncertainty refers to the residuals, and model uncertainty refers to the standard errors of the estimated coefficients.

To characterize data uncertainty, either a parametric or non-parametric method can be used. Parametric methods refer to the models that parameterize a probabilistic distribution. Parametric models are often estimated using Bayesian methods or mean-variance estimation (MVE). Although conceptually appealing, Bayesian methods often involve unnecessarily intense computation, which relies on sampling methods and variational inference [117, 76]. MVE minimizes the negative log-likelihood (NLL) loss based on a pre-specified distribution of the dependent variable [111, 75]. The MVE is computationally efficient but could lead to misleading results if probabilistic distributions are misspecified. Non-parametric methods quantify uncertainty without explicitly imposing any parametric form. Typical examples include quantile regression [80] and Lower Upper Bound Estimation (LUBE) [117, 76]. Non-parametric methods do not have misspecification problems, but optimizing in complex neural networks is hard. The pros and cons of both methods are systematically compared in review articles [70, 41].

Besides data uncertainty, uncertainty also emerges in model construction and training. Bayesian deep neural networks refer to the models that characterize all parameters of neural networks by probability distributions, typically Gaussian distribution [194, 17]. A popular, low-cost way to do approximate Bayesian inference for neural networks with Gaussian parameters is to use dropouts [89, 39, 136]. Another source of model uncertainty comes from model training. Neural networks tend to converge to different local minima, resulting in different predictions. This type of uncertainty is usually addressed by bootstrapping or model ensembles [55, 83]. Bootstrapping refers to the process of repeated training with re-sampling to capture the distributional uncertainty in sampling. Model ensembling refers to an average of multiple training procedures with different parameter initializations.

Although the literature on travel demand uncertainty within the deep learning framework is scarce, classical statistics have decades of effort in quantifying

travel demand uncertainty, focusing on evaluating a rich family of probabilistic assumptions. Using high-resolution temporal ridership data, researchers adopted the ARIMA-GARCH model to quantify the volatility of subway demands during special events [25] and adaptive Kalman filters to analyze real-time transit ridership [47]. Using cross-sectional data, researchers used bootstrapping [118] to analyze parameter uncertainty, ensembling [27] to analyze activity uncertainty, and heteroskedastic errors to describe the association between social demographics and ridership uncertainty in discrete choice models [118]. The common debates regarding the probabilistic assumptions include homoskedastic vs. heteroskedastic errors, Gaussian vs. exponential tails, real-line vs. positive support, and many others [25, 147, 188], through which researchers could learn the detailed characteristics of travel demand uncertainty. The primary focus of classical statistics on probabilistic assumptions presents an interesting difference from the primary focus of deep learning in refining deterministic assumptions. Indeed, it is quite ambiguous which set of assumptions is more crucial in quantifying travel demand uncertainty.

2.3 Theoretical Framework

A probabilistic graph convolutional neural network is proposed to compare the deterministic and probabilistic assumptions. The probabilistic GNN is represented as:

$$y \sim \mathcal{G}(\theta) = \mathcal{G}(\mathcal{F}(\mathbf{X}, w)) \quad (2.1)$$

The first statement is a probability assumption about y , specified by the model family $\mathcal{G}(\theta)$, and the second uses the model family $\mathcal{F}(\mathbf{X}, w)$ to parameterize the probability distributions of $\mathcal{G}(\theta)$, with \mathbf{X} being the inputs and w being the model weights. While the recent studies focus on enriching $\mathcal{F}(\mathbf{X}, w)$ through spatiotemporal deep learning architectures, the potentially more critical probability assumption $\mathcal{G}(\theta)$ is largely neglected. We compare the effects of the probabilistic assumptions \mathcal{G} and the deterministic assumptions \mathcal{F} in determining the model performance.

Table 2.1 summarizes the specifics of probabilistic and deterministic assumptions that substantiate the probabilistic GNN framework. The upper panel lists the probabilistic assumptions and their probability density functions. The bottom panel lists the deterministic architecture and hyperparameters. The six probabilistic assumptions tested are Homoskedastic Gaussian (HomoG), Poisson (Pois), Heteroskedastic Gaussian (HetG), Truncated Gaussian (TG), Gaussian Ensemble (GEns), and Laplace (LAP). Two deterministic architectures, Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT), are used to specify the probabilistic parameters. A cross-product of the two panels is used to substantiate the probabilistic GCNs, leading to twelve base models: HomoG-GCN, HomoG-GAT, HetG-GCN, HetG-GAT, TG-GCN, TG-GAT, GEns-GCN, GEns-GAT, Pois-GCN, Pois-GAT, Lap-GCN, and Lap-GAT.

Table 2.1: Model Design of the Probabilistic Graph Neural Networks Framework

Panel 1: Probabilistic Assumptions in \mathcal{G}		
Probabilistic assumptions	Probability density function	Distribution parameters θ
Homoskedastic Gaussian (HomoG)	$f_{HomoG}(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}), \sigma = c$	μ
Poisson (Pois)	$f(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$	λ
Heteroskedastic Gaussian (HetG)	$f_{HetG}(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2})$	μ, σ
Truncated Gaussian (TG)	$f_{TG}(x; \mu, \sigma) = \frac{f_G(x; \mu, \sigma)}{1 - f_G(0; \mu, \sigma)}$	μ, σ
Gaussian Ensemble (GEns)	$y_* \sim \mathcal{N}(\frac{1}{K} \sum_k \mu_k, \frac{1}{K} \sum_k (\sigma_k^2 + \mu_k^2) - \mu_*^2)$	μ_k, σ_k for $k = 1..K$ models
Laplace (Lap)	$f(x; \mu, b) = \frac{1}{2b} \exp(-\frac{ x-\mu }{b})$	μ, b
Panel 2: Deterministic Assumptions in \mathcal{F}		
Deterministic assumptions	Graph convolutional iteration function	Deterministic parameters w
Graph Convolutional Networks (GCN) [77]	$h^{(l+1)} = \sigma(\sum_{r=1}^R \tilde{D}_r^{-\frac{1}{2}} \tilde{A}_r \tilde{D}_r^{-\frac{1}{2}} h^{(l)} W_r^{(l)})$	$W_r^{(l)}$
Graph Attention Networks (GAT) [145]	$h^{(l+1)} = \sigma(\sum_{j \in N(i)} \alpha_{ij} W h_j^{(l)}), \alpha_{ij} = \frac{\exp(W_i h_j^{(l)})}{\sum_{k \in N_i} \exp(W_i h_k^{(l)})}$	α_{ij}, W
Other hyperparameters in deterministic assumptions \mathcal{F} : Number of GCN/GAT/LSTM layers, Number of hidden layer neurons, Dropouts, Weight decay.		

2.3.1 Probabilistic Assumptions \mathcal{G}

The Gaussian distribution with a deterministic and homoskedastic variance term is chosen as the benchmark in specifying \mathcal{G} because it is stable and has simple ensembling properties. This homoskedastic Gaussian assumption represents many deterministic deep learning models that use mean squared error as the training objective. The Gaussian benchmark facilitates comparing the probabilistic assumptions, as listed below.

1. Homoskedasticity vs. heteroskedasticity

We compare the heteroskedastic and homoskedastic Gaussian assumptions to examine how the data variance influences the model performance. The homoskedastic assumption assumes the same variance for all observations, whereas the heteroskedastic assumption estimates variance for every observation. The travel demand variances likely have spatiotemporal patterns.

2. Continuous vs. discrete support

We compare the Poisson distribution to the homoskedastic Gaussian benchmark to examine the effectiveness of continuous vs. discrete supports in determining the model performance. Since ridership takes integer values, the Poisson distribution could be more appropriate. However, the Poisson distribution uses only one parameter to represent both mean and variance, which could be overly restrictive.

3. Real-line vs. non-negative support

We compare the truncated heteroskedastic Gaussian distribution to the heteroskedastic Gaussian distribution to examine the effectiveness of the real-line vs. non-negative distribution support. Travel demand is non-negative, but the support of the Gaussian distribution covers the entire real line $(-\infty, +\infty)$. Therefore, a normal distribution left-truncated at zero is implemented to test whether non-negativity should be strictly imposed.

4. Gaussian vs. exponential tails

We compare the Laplace distribution to the heteroskedastic Gaussian distribution to examine whether the tail behavior of the distribution matters. The probability density function of the Gaussian distribution decays at a fast rate of e^{-x^2} , while the Laplace distribution has a heavier tail with the decay rate of $e^{-|x|}$.

5. Single vs. ensembled models

We compare the ensembled heteroskedastic Gaussian distributions to a single heteroskedastic Gaussian model to test whether an ensemble of distributions can outperform a single distributional assumption. The ensemble model is created by uniformly mixing K estimates, which are trained independently with different parameter initializations. Suppose $\mathcal{Y}_k, \sigma_k^2$ are the estimated mean and variance for model k , and \mathcal{Y}_* is the ensembled random variable. For the Gaussian distribution, the mixture can be further approximated as a Gaussian distribution, whose mean and variance are, respectively, the mean and variance of the mixture. The ensembled distribution is given by [83]: $\mathcal{Y}_* \sim \mathcal{N}(\frac{1}{K} \sum_k \mathcal{Y}_k, \frac{1}{K} \sum_k (\sigma_k^2 + \mathcal{Y}_k^2) - \mathcal{Y}_*^2)$.

With the distributional assumption, maximum likelihood estimation (MLE) is used to learn the parameters, which translates to minimizing the negative log-likelihood (NLL) loss in implementation. The NLL loss function based on the joint density of all observations is simply the negative sum of the log of the probability density of all observations:

$$\text{NLL} = - \sum_{s,t} \log \mathcal{G}(y_{st} | \mathbf{X}_t; w) \quad (2.2)$$

In the uncertainty quantification literature, researchers typically differentiate between data (aleatoric) and model (epistemic) uncertainty [78], and their sum is referred to as prediction uncertainty. The model uncertainty refers to the uncertainty resulting from the difference between $\mathcal{F}(\mathbf{X}, w)$ and the estimated $\hat{\mathcal{F}}(\mathbf{X}, w)$. The data uncertainty refers to the randomness in the data generation process and is represented by probabilistic assumptions \mathcal{G} , of which y has the distribution. The prediction uncertainty combines model and data uncertainty and describes the overall difference

between the actual y and the predicted \hat{y} . The relationship between the three quantities is directly given by $\sigma_y^2 = \sigma_{model}^2 + \sigma_{data}^2$. In our framework, assumptions 1-4 deal with the data uncertainty alone, while assumption 5 quantifies the prediction uncertainty.

2.3.2 Deterministic Assumptions in \mathcal{F}

The deterministic assumptions in \mathcal{F} consist of the spatial encoding, temporal encoding, and associated hyperparameter specifications. The formulation of common spatial and temporal encodings is discussed below.

GCN and GAT for Spatial Encoding

Two predominant spatial encoding methods - GCN and GAT - are adopted and compared to examine the effect of the deterministic architectures on model performance. The GCN layers need to access the global structure of the graph in the form of adjacency matrix(es), while the GAT layers aim to learn the spatial correlations from data. The propagation formula of both layers is introduced in Table 2.1.

To construct the multi-graph proximity in GCNs and GATs, four types of adjacency matrices are computed, including direct connectivity $[A_{Con}]_{ij} = 1$ if two stations are adjacent, $= 0$ otherwise, network distance $[A_{Net}]_{ij} = \text{Network Distance}(i, j)^{-1}$, Euclidean distance $[A_{Euc}]_{ij} = \text{Euclidean Distance}(i, j)^{-1}$, and functional similarity $[A_{Func}]_{ij} = \sqrt{(F_i - F_j)^T (F_i - F_j)^{-1}}$, where F_i is the vector of functionalities represented by population, jobs, percent low income, percent minority, percent adults, number of schools, shops, restaurants, etc.

LSTM for Temporal Encoding

The LSTM network is chosen for temporal encoding. In short, the LSTM layers take the spatially encoded tensors for l previous time periods as inputs and propagate them down the layers through a series of input, forget, cell state (memory), and output gates. To make the prediction, the encoded output from the last time step

h_{t-1} is decoded by fully connected layers to get the contribution (weights) of the time series on the final prediction. The LSTM structure is well-documented in many papers [126, 25], and will not be discussed in detail here.

2.3.3 Specific Examples: Gaussian Distributions and Mean-Variance Estimation

As an important case, the Gaussian distribution with homoskedastic variance (HomoG-GCN and HomoG-GAT) represents the vast number of deterministic deep learning models that use mean squared errors as the training objective. With the homoskedastic Gaussian assumption, the dependent variable y follows Gaussian distribution with mean $F_1(\mathbf{X}, w_1)$ and a constant variance c . The MLE with this homoskedastic Gaussian distribution is the same as the deterministic deep learning with the mean squared errors as the training objective because

$$\log P(x; \mu, \sigma = c) = \log \frac{1}{c\sqrt{2\pi}} - \frac{1}{2} \frac{(x - \mu)^2}{c^2} \quad (2.3)$$

In other words, our benchmark example represents the predominant deterministic modeling technique in this field.

The homoskedastic Gaussian can be extended to the heteroskedastic Gaussian distribution, which resembles the mean-variance estimation (MVE), the most dominant method in uncertainty quantification literature. With the heteroskedastic Gaussian distribution, $y \sim \mathcal{N}(\theta) = \mathcal{N}(F_1(\mathbf{X}, w_1), F_2(\mathbf{X}, w_2)) = F_1(\mathbf{X}, w_1) + \mathcal{N}(0, F_2(\mathbf{X}, w_2))$. Essentially, the MVE uses two graph neural networks to capture the mean and variance separately, which is the same as the HetG-GCN and HetG-GAT models. Therefore, the two Gaussian examples in our probabilistic GNN framework can represent the two most common research methods: deterministic spatiotemporal models and MVE for uncertainty quantification.

2.3.4 Evaluation

Performance metrics can be grouped into three categories: composite measures, point prediction quality, and uncertainty prediction quality. Table 2.2 summarizes the evaluation metrics. The NLL loss is a composite metric to evaluate the joint quality of point and uncertainty estimates. The standard mean absolute error (MAE) and mean absolute percent error (MAPE) are used to evaluate point predictions.

Table 2.2: Three Categories of Evaluation Metrics

Category	Metric	Formula
Composite	Negative Log Likelihood	$NLL = -\sum_i \log P_W(y_i \mathbf{X}_i)$
Point	Mean Absolute Error Mean Absolute Percent Error	$MAE = \frac{1}{N} \sum_i y_i - \hat{y}_i $ $MAPE = MAE/\bar{y}$
Uncertainty	Calibration Error Mean PI Width PI Coverage Probability	$CE = \sum_{p=0}^1 q(p) - p $ $MPIW = \frac{1}{n} \sum_{i=1}^N (U_i - L_i)$ $PICP = \frac{1}{N} \sum_{i=1}^n \mathbf{1}\{L_i \leq y_i \leq U_i\}$

The quality of the uncertainty estimates is less straightforward to evaluate because the ground truth distributions are unknown. Therefore, we designed the calibration error (CE) metric and visualized quantile-quantile plots to measure the distributional fit. Let $F_i(y)$ represent the cumulative distribution function of observation i . Define $q(p) = \mathbb{P}(F_i(y_i) \leq p)$ as the proportion of observations that actually fall into the p -th quantile of the estimated distribution. A well-calibrated model should generate distributions that align with the empirical distribution so that $q(p) = p$. Therefore, the calibration error associated with the quantile-quantile plots is defined as the sum of the deviation between the empirical quantiles and the predicted quantiles, approximated by several discrete bins from $[0, 1]$: $CE = \sum_{p=0}^1 |q(p) - p|$. An alternative metric is the simultaneous use of the Prediction Interval Coverage Probability (PICP) and the Mean Prediction Interval Width (MPIW) [117, 76, 70]. Formally, $PICP = \frac{1}{n} \sum_{i=1}^n c_i$, where $c_i = \mathbf{1}\{L_i \leq y_i \leq U_i\}$, that is an indicator variable of whether observation i falls within the prediction interval bounded by L_i and U_i . $MPIW = \frac{1}{n} \sum_{i=1}^n (U_i - L_i)$ measures the average width of the intervals. With significance level $1 - \alpha$, the lower

bound of the prediction interval is given by $L = F_i^{-1}(\frac{\alpha}{2})$, and the upper bound of the prediction interval by $U = F_i^{-1}(1 - \frac{\alpha}{2})$, where F_i^{-1} is the predicted inverse cumulative function of observation i . Using this approach, evaluating uncertainty quantification involves a tradeoff between PICP and MPIW. A high-quality prediction interval should be narrow while covering a large proportion of data points; however, wider MPIW naturally leads to larger PICP. Therefore, this tradeoff poses a challenge in model selection since it is difficult for one model to dominate in both metrics.

2.4 Case Study

Two case studies were conducted to examine the effect of deterministic and probabilistic assumptions on predicting travel demand, focusing on estimating uncertainty. The case studies use data from the Chicago Transit Authority’s (CTA) rail system and ridesharing in Chicago.

This section describes data sources, the spatial and temporal characteristics of the two datasets, and the experiment setup. Our experiments are implemented in PyTorch, and the source code is available at <https://github.com/sunnyqywang/uncertainty>.

2.4.1 Data

The CTA rail and bus data was obtained through collaboration with the CTA, while the ridesharing data was sourced from the City of Chicago Open Data Portal¹. The CTA rail system has tap-in records for each trip, which were aggregated into 15-minute intervals for temporal analysis. The system comprises 141 stations arranged in a graph structure for spatial analysis. The spatial relationships between stations can be identified by adjacency matrices constructed from Euclidean, connectivity, network, and functional similarity. Ridesharing trips are available at the census tract level in 15-minute intervals. The graph structure is determined by the relationships

¹<https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p>

between census tracts. As there is no explicit network, the ridesharing adjacency matrices are defined by Euclidean, connectivity (neighboring census tracts are considered connected), and functional similarity only. Most census tracts have close to no ridesharing trips in most 15-minute intervals. Since learning sparse graphs is a topic of its own [182], the ridesharing dataset is filtered to include only those census tracts that, on average, had more than 30 trips per hour pre-COVID. This resulted in a total of 59 census tracts being used in the analysis.

Moran’s I was calculated for each 15-minute ridership snapshot using each of the weight matrices to demonstrate that spatial autocorrelation exists through the defined adjacency matrices. Figure 2-1 shows the histograms of Moran’s I for both data sources. Most statistics have averaged well above 0, indicating the existence of spatial autocorrelations defined by the adjacency matrices. Ridesharing trips appear to correlate less with Euclidean distance and functional similarity, but the correlation between bordering tracts is very strong.

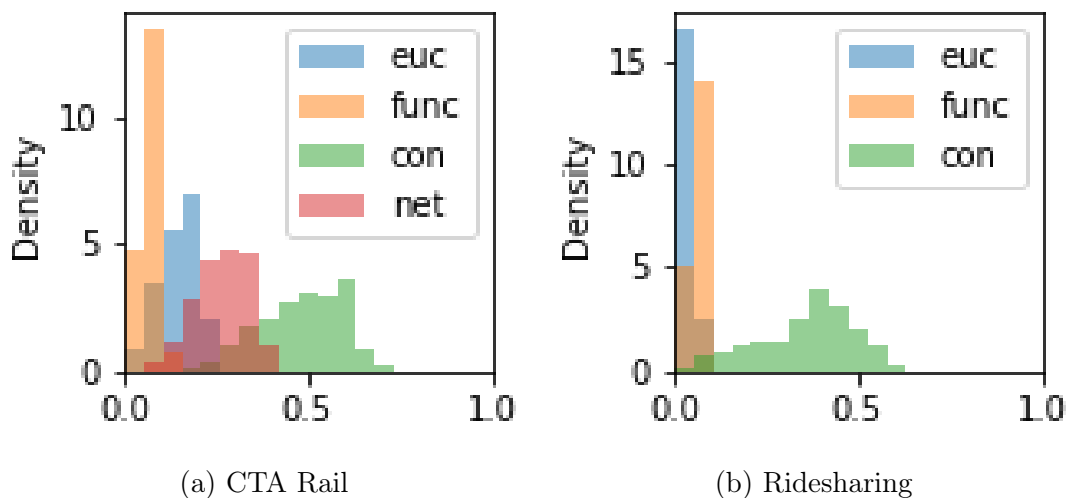


Figure 2-1: Spatial Autocorrelation (Moran’s I) of CTA Rail and Ridesharing Data.

Five different explanatory variables are obtained for each data source. First, the system’s history, recent (in the past 1.5 hours) and historical (last week, same time period), is used to account for the temporal correlation. Second, the history of the other travel modes is included to mine the correlation between different modes. In the case of CTA rail, ridesharing and bus counts are included, and in the case of

ridesharing, CTA rail and bus counts are included. Third, demographics² and POIs³ are spatial covariates used in the calculation of functional similarity. Fourth, weather⁴ is a temporal covariate that is assumed to be the same in the whole study area. Lastly, the frequency of services in each spatial unit during each period indicates supply levels. In the case of ridesharing, we do not have supply information. Instead, we use the bus frequency as a proxy. Bus schedules are slower to respond during normal times, but after March 2020, the CTA is actively adjusting service to account for labor shortages and changing demand.

2.4.2 Experiment Setup

The datasets were split into three subsets along the temporal dimension: train, validation, and test. The training set consists of data from August 1, 2019, to Feb 16, 2020; the validation set from Feb 17, 2020, to March 1, 2020; and the testing set consists of four different two-week periods during the course of COVID-19 pandemic in 2020: immediately before (March 2 to March 15), stay-at-home (March 16 to March 29), initial recovery (June 22 to July 5), and steady recovery (Oct 18 to Oct 31). Significant changes emerged starting in March 2020 when the confirmed COVID-19 cases were growing rapidly, and the city issued stay-at-home orders. Since then, the ridership has seen different stages of recovery. The changes in ridership of a few stations/census tracts are shown in Figure 2-2. Meanwhile, changes induced by COVID-19 provide an opportunity to test the temporal generalizability from pre- to post-COVID periods.

Several other architectural considerations have been considered besides the GCN/GAT spatial convolutions. Figure 2-3 summarizes the deterministic architecture, consisting of three components: spatiotemporal layers (GCN/GAT+LSTM) for recent observed demand, and linear layer to connect last week’s observation (**history**), and weather (temperature and **precipitation**). The three components are indexed by r, h, p , respec-

²<https://www.census.gov/programs-surveys/acs/data.html>

³<https://planet.openstreetmap.org/>

⁴<https://www.ncdc.noaa.gov/cdo-web/>

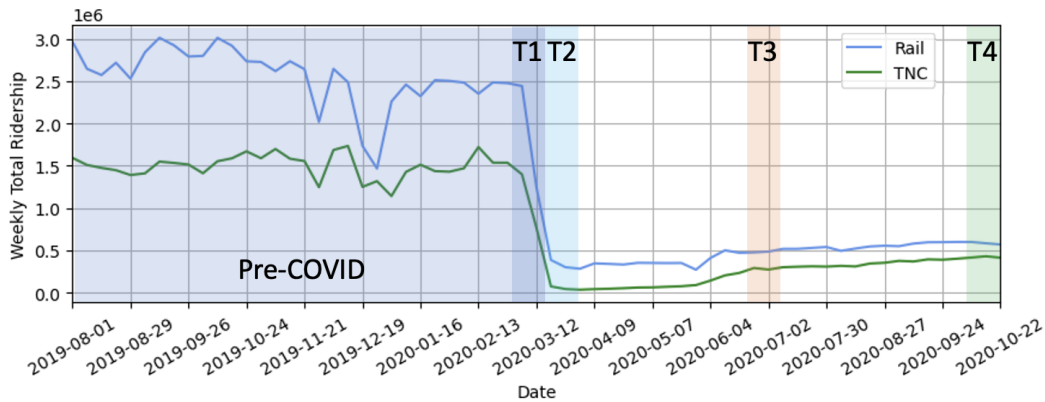


Figure 2-2: Average Daily Ridership of CTA Rail and Ridesharing (T1: Immediately Before; T2: Stay-at-home; T3: Initial Recovery; T4: Steady Recovery).

tively, and summed to the final prediction. First, due to the cyclic nature of travel demand, the travel demand time series can be decomposed into a weekly component $(\hat{\mathcal{Y}}_t^h, \hat{\sigma}_t^h)$ and a time-of-day component $(\hat{\mathcal{Y}}_t^r, \hat{\sigma}_t^r)$. The weekly component is calculated from a reference demand. A good reference demand is the observed value for the same region and time in the previous week. The weights w_h are obtained from the LSTM encodings. Additionally, the recent demand for LSTM network inputs is the deviation from last week’s demand, and the decoded outputs are used to produce both the weekly and the time-of-day components. Intuitively, if the recent residual demand is very different from the reference, the reference should have a smaller weight on the final prediction, which is more uncertain. The time-of-day component is directly obtained from LSTM encodings of the recent residual demand. Next, weather is another source of temporal variation because extreme weather could have an evident impact on transit ridership. The model takes daily deviations from average of precipitation P_T and temperature T_E , and multiply them with spatiotemporal weights $W^p, W^t \in \mathbb{R}^{2 \times T \times S}$ to get $[\hat{\mathcal{Y}}_t^p, \hat{\sigma}_t^p]$.

For each dataset and each deterministic architecture, six probabilistic assumptions are tested: homoskedastic Gaussian (HomoG), Poisson (Pois), heteroskedastic Gaussian (HetG), truncated Gaussian (TG), Laplace (Lap), ensembled Gaussian (GEns). In the HomoG model, a search for the best standard deviation was done.

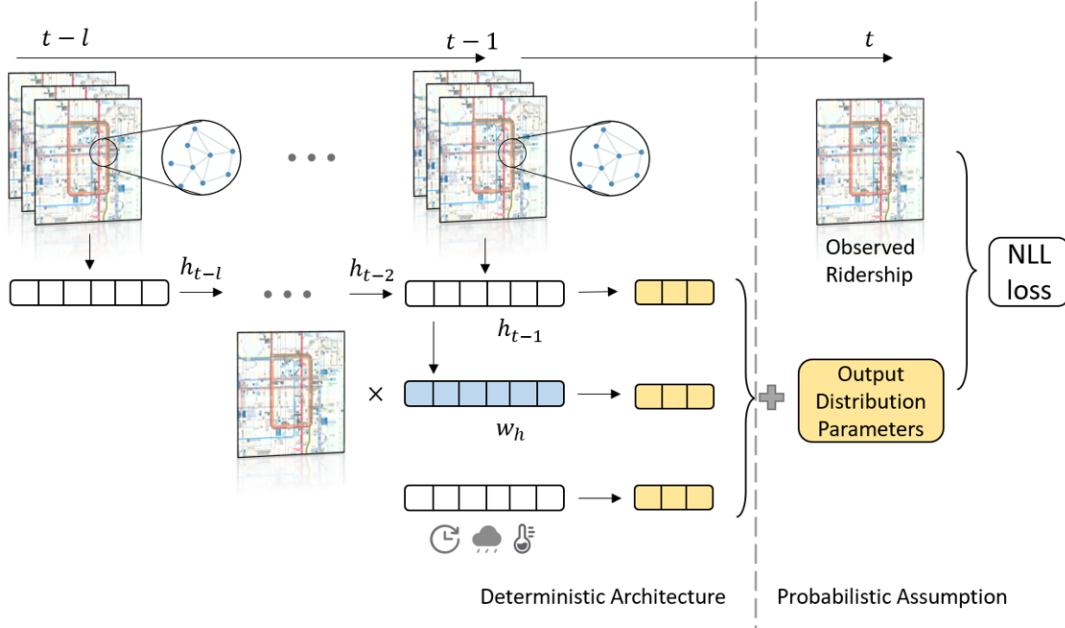


Figure 2-3: Proposed model architecture of Probabilistic Graph Neural Networks.

Values are searched in multiples of \bar{y} : $\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$, and the best value for the variance is $\frac{1}{2}\bar{y}$ for both datasets. In GEnS, five top HetG models by validation set loss are selected to create an ensemble model. The ensemble distribution is given by [83]: $\mathcal{Y}_* \sim \mathcal{N}(\frac{1}{K} \sum_k \mathcal{Y}_k, \frac{1}{K} \sum_k (\sigma_k^2 + \mathcal{Y}_k^2) - \mathcal{Y}_*^2)$.

2.5 Results and Discussion

We present and compare the model performances of the Prob-GNNs under six probabilistic and two deterministic assumptions on two separate datasets. First, we analyze the model performance for periods immediately after the training period, from March 2 to March 15, 2020. We demonstrate that probabilistic assumptions can influence model performance more significantly than deterministic architectures. Next, we show that uncertainty predictions are important, especially when point predictions become unreliable during system disruptions, by applying models trained with the pre-COVID training set to the post-COVID test sets. Lastly, we discuss the spatiotemporal patterns of uncertainty revealed by the models.

First, to illustrate the models' proper convergence, Figure 2-4 presents the GCN

model training curves for all probabilistic assumptions trained on CTA Rail. All models demonstrated successful convergence under the same learning rate. However, the number of epochs required for convergence varied among the models. The two-parameter distributions (HetG, TG, Laplace) converged faster, achieving a lower NLL than the one-parameter distributions (HomoG, Pois).

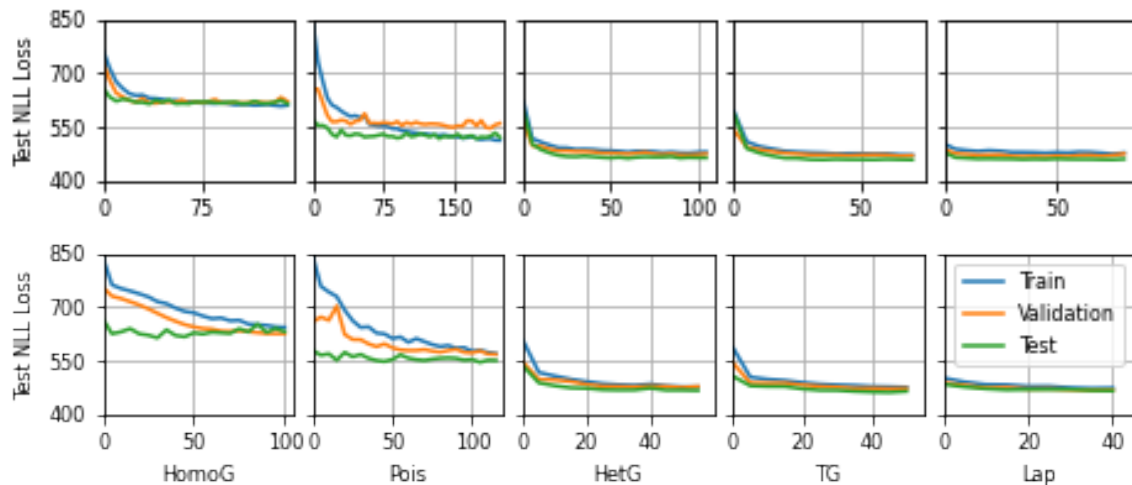


Figure 2-4: Training Curves. Top: GCN; Bottom: GAT

Model performances of CTA Rail and ridesharing data on the test set between March 2 and March 15, 2020, are tabulated in Tables 2.3. The table has two panels for GCN and GAT models, respectively. Each panel presents the six probabilistic assumptions in Section 2.4.2. Model performances are measured on three categories of metrics: composite (NLL), uncertainty (calibration error, MPIW, and PICP), and point (MAE, MAPE). In the tables, the intended level of coverage was set to 95%. The next two sections discuss the findings from Table 2.3 in detail.

2.5.1 Significance of Probabilistic and Deterministic Assumptions

The quality of uncertainty quantification varies remarkably with probabilistic assumptions. The performance is significantly higher in the probabilistic models with two parameters (HetG and Lap) compared to the single-parameter models (HomoG

Table 2.3: Prob-GNN Model Performance (Test Period: Immediately Before - March 2 to March 15, 2020)

Model	CTA Rail						Ridesharing					
	Comp	Uncertainty Prediction			Point Prediction		Comp	Uncertainty Prediction			Point Prediction	
	NLL	Cal. Err	MPIW	PICP	MAE	MAPE	NLL	Cal. Err	MPIW	PICP	MAE	MAPE
HomoG-GCN	612.5	0.146	64.53	97.9%	7.64	18.5%	225.6	0.114	42.63	97.7%	5.81	25.2%
Pois-GCN	529.2	0.075	18.04	84.9%	7.18	17.4%	211.9	0.083	15.53	82.7%	5.81	25.2%
HetG-GCN	462.9	0.049	39.04	96.5%	7.67	18.6%	187.6	0.019	27.32	93.3%	5.80	25.1%
TG-GCN	480.7	0.021	39.37	95.6%	7.61	18.5%	196.7	0.029	30.63	95.2%	6.06	26.3%
Lap-GCN	460.2	0.026	43.48	97.0%	7.52	18.2%	187.1	0.022	34.12	96.5%	5.81	25.2%
GEEns-GCN	459.6	0.036	37.48	95.7%	7.42	18.0%	185.6	0.017	28.01	93.9%	5.70	24.7%
HomoG-GAT	620.4	0.141	64.54	97.8%	8.13	19.7%	230.7	0.112	43.26	96.9%	6.52	28.2%
Pois-GAT	572.2	0.100	18.11	81.8%	7.98	19.4%	214.9	0.093	15.62	82.0%	6.06	26.3%
HetG-GAT	472.3	0.058	39.58	96.3%	7.84	19.0%	188.3	0.041	31.43	95.7%	6.07	26.3%
TG-GAT	481.6	0.026	40.96	96.6%	7.69	18.7%	199.9	0.024	32.54	95.7%	6.55	28.4%
Lap-GAT	464.3	0.037	42.77	97.7%	7.63	18.5%	189.5	0.021	36.08	96.5%	6.16	26.7%
GEEns-GAT	470.5	0.037	42.49	96.8%	7.52	18.3%	188.1	0.044	32.28	96.0%	6.18	26.8%

and Pois). Among all GCN models, the average NLL and calibration error for two-parameter distributions in the CTA Rail data is 465.9 and 0.033, and for one-parameter distributions, 570.8(+22.5%) and 0.11(3.3 \times). Similar observations can be made in the ridesharing dataset, where the average NLL is 218.8 vs. 189.3(+15.4%) and the calibration error is 0.017 vs. 0.099(5.8 \times). Similar observations can be made for GAT models.

Truncated Gaussian and Laplace distributions have the overall best distributional fit. Figure 2-5 illustrates the distributional fit by plotting the empirical quantiles $q(p)$ against the theoretical quantiles p for both datasets across all the GCN models. The line $y = x$ represents a perfectly-calibrated model. The calibration error in Table 2.3 corresponds to the areas between the $y = x$ line and the empirical quantiles. Poisson and homoskedastic Gaussian are very far from the $y = x$ line, while truncated Gaussian and Laplace trace the line most closely. Although the Heteroskedastic Gaussian works the best with GCN models on ridesharing data, the truncated Gaussian and Laplace distribution had more stable performances across the datasets and models, and the curves are closer to the $y = x$ line.

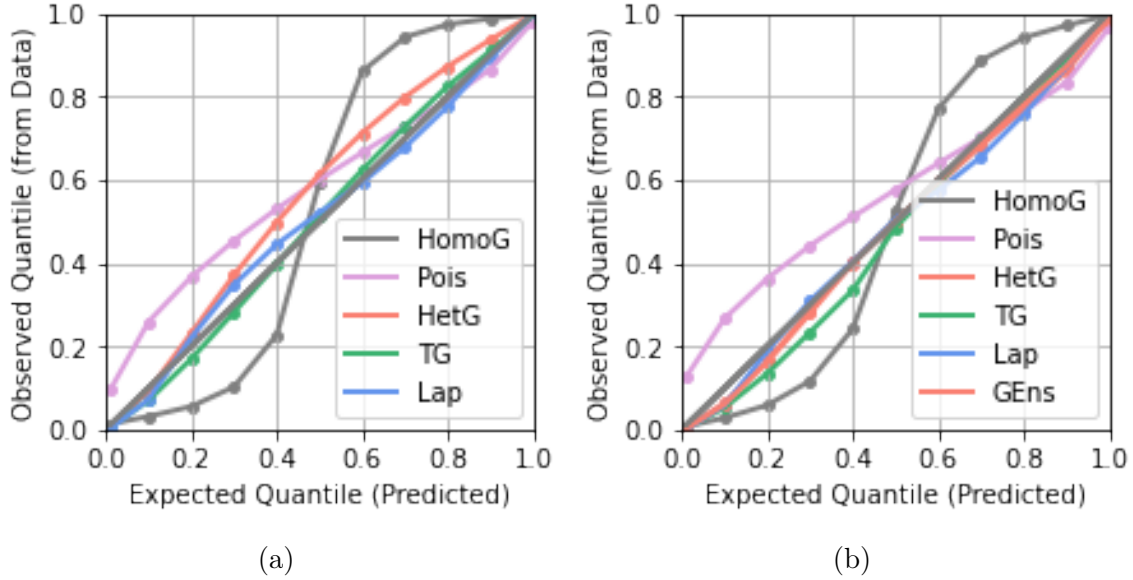


Figure 2-5: Calibration Plot of GCN Models (a) CTA rail and (b) Ridesharing.

In contrast to the strong influence probabilistic assumptions have on the quality of uncertainty predictions, little influence is exerted on the point prediction quality from the probabilistic assumptions. For both datasets and for both GCN and GAT, between different probabilistic assumptions, the performance gap between the best and the worst point estimate is around 4%, significantly less than that of the composite and the uncertainty metrics.

The variations in predictive performance caused by the deterministic assumptions are also small compared to the probabilistic assumptions. Comparing the GCNs and GATs, the error metrics are similar, and the performance patterns across probabilistic assumptions remain the same. The one-parameter probabilistic assumptions are likely to be affected more by deterministic architectures. For all probabilistic assumptions except for Poisson, less than a 3% difference in NLL loss is observed between GCN and GAT models for the same dataset. Although the GAT allows the model to learn spatial relationships and is more flexible, the predefined adjacency matrices serve as domain knowledge to reduce the learning complexity. Theoretically, the GAT setup should perform better with more complex relationships or larger sample sizes, and the GCN setup is better for more efficient learning under limited sample sizes. In this case, there is not a large difference between the two architectures.

2.5.2 Implications of Probabilistic Assumptions

The probabilistic assumptions not only suggest the distributional fit but also have practical implications on the ridership patterns. We then discuss these implications in detail below.

Heteroskedasticity vs. Homoskedasticity (HetG vs. HomoG)

HetG forces a constant variance across all observations, resulting in an inaccurate data representation. The HomoG models achieve around 20% higher NLL loss and 1.5 to 2 times higher calibration error than the HetG models. For example, the NLL loss of HetG-GCN is 462.9, 24% lower than the NLL loss (612.5) of HomoG-GCN for CTA Rail. Although predicting the same variance is inaccurate distributionally, the average MPIW and PICP are not much worse than their heteroskedastic counterparts since the variance scale is searched and fixed. Additionally, the inaccurate distributional assumption does not affect the quality of the point estimate, and point estimate errors from HomoG models are only 1-2% higher than the best model, as the likelihood improvement can only come from a more accurate prediction of the mean. Regardless, having one fixed uncertainty parameter is an inaccurate representation and limits the model's flexibility to adapt to sharp changes in ridership magnitude.

Continuous vs. Discrete (HomoG vs. Pois)

Both HomoG and Pois have the worst NLL loss among all probabilistic assumptions tested. Despite its discreteness, the Poisson assumption yields a similar or worse NLL and calibration error than HomoG since it forces equal mean and variance. However, for both datasets, the prediction intervals significantly under-predict the demand, meaning that the variance should be larger than the mean. The PICP of the Poisson prediction interval indicates the magnitude of variance compared to the mean. The smaller the PICP, the larger the variance.

Real-line vs. non-negative support (HetG vs. TG)

Since the time resolution is 15 minutes, left-truncation at 0 significantly improves the model performance as many predictions have means close to 0. TG is the best-performing model for the CTA Rail and the second-best for ridesharing. Even though the average ridership is around 40 per station and 23 per census tract, significant heterogeneity exists among different stations/census tracts. Sparsity (zero ridership) is an issue to be considered in short-term or real-time demand predictions.

Gaussian vs. Exponential tails (HetG vs. Lap)

Both distributions are characterized by two parameters and can be trained to describe the data accurately. The weights of the tails are different, and the model performances on the two probabilistic assumptions suggest behavioral differences. Compared to HetG, Lap has a heavier tail; hence, the prediction intervals tend to be larger, covering more observations in more extreme cases. In all cases, Lap covers more points than intended.

Single vs. ensembled models (HetG vs. GEns)

Ensembling only slightly improves the model results. The NLL loss, calibration error, and MAE improve by less than 1% in all cases. For example, the biggest improvement occurs in the GAT models for CTA Rail data, with NLL loss for HetG-GAT being 462.9 and for GEns-GAT 459.6. Since ensembling aims to reduce model uncertainty, its ineffectiveness suggests that different training instances of the neural networks produce similar results, and the model uncertainty is low.

We also constructed prediction intervals with Monte Carlo dropouts to further illustrate that model uncertainty is much smaller than the data uncertainty inherent in the data generation process. Monte Carlo dropout approximates Bayesian neural networks and measures the model uncertainty by applying dropout at test time, treating each as an instance from the space of all available models. However, since the different training instances produce similar results, Monte Carlo dropout fails

to capture the full picture. Its prediction intervals were very narrow, with PICPs between 30% - 40% for both datasets. Additionally, since we applied dropout at test time, the point prediction loses the benefit of dropout regularization and performs worse than other models.

2.5.3 Model Performance under System Disruption

The drastic change in pre- and post-COVID periods presents a unique opportunity to test the generalizability of Prob-GNNs. The previous sections use the periods immediately following the training set for validation and testing. This section compares the predictive performance at different stages of the COVID-19 pandemic by applying the models trained with the pre-COVID training set to three post-COVID time periods. Table 2.4 presents CTA Rail and ridesharing results. Since HomoG and Pois performed poorly in the previous test set, they are excluded from this comparison.

Table 2.4: Model Generalization under System Disruption

Model	CTA Rail						Ridesharing					
	Comp	Uncertainty Prediction			Point Prediction		Comp	Uncertainty Prediction			Point Prediction	
	NLL	Cal. Err	MPIW	PICP	MAE	MAPE	NLL	Cal. Err	MPIW	PICP	MAE	MAPE
Stay-at-home (March 16 - March 29, 2020)												
HetG-GCN	397.9	0.258	16.44	96.0%	4.92	105%	165.2	0.291	11.89	88.1%	4.40	198%
TG-GCN	388.9	0.176	16.54	95.2%	4.38	93.6%	163.4	0.114	13.04	97.1%	2.96	133%
Lap-GCN	405.8	0.247	17.65	98.7%	4.34	92.9%	148.3	0.217	13.41	96.7%	3.70	167%
GEs-GCN	400.4	0.263	16.14	95.6%	4.89	105%	162.4	0.286	11.94	89.9%	4.20	189%
Initial Recovery (June 22 - July 5, 2020)												
HetG-GCN	413.7	0.214	18.19	94.9%	4.46	61.6%	161.9	0.171	12.39	90.2%	3.57	68.3%
TG-GCN	436.2	0.174	18.29	94.3%	4.37	60.3%	173.3	0.037	12.91	93.8%	2.88	55.1%
Lap-GCN	409.4	0.190	18.83	98.3%	3.95	54.5%	161.7	0.071	13.81	94.2%	2.82	54.0%
GEs-GCN	412.4	0.213	18.45	95.0%	4.4	60.7%	152.2	0.112	11.22	92.0%	2.88	55.1%
Steady Recovery (Oct 12 - Oct 25, 2020)												
HetG-GCN	387.6	0.017	16.79	93.7%	3.38	36.2%	174.7	0.128	15.80	89.1%	4.33	51.4%
TG-GCN	389.5	0.030	18.46	95.9%	3.90	41.7%	186.0	0.035	17.16	92.9%	4.03	47.8%
Lap-GCN	406.6	0.057	21.63	96.1%	4.12	44.1%	170.6	0.073	18.62	94.4%	3.93	46.7%
GEs-GCN	391.1	0.053	20.28	95.1%	3.44	36.8%	172.7	0.098	14.75	89.2%	3.95	47.0%

All the error metrics increase under the significant domain shifts from pre- to post-

COVID. In the stay-at-home period, the average ridership for both systems dropped to less than 10% of pre-COVID levels. The NLL, MPIW, and MAE are not comparable to pre-COVID levels because they are influenced by the magnitude of ridership, while calibration error, PICP, and MAPE are unit-free and can be compared to values in Table 2.3. Unsurprisingly, the performance during the stay-at-home period is relatively poor but slowly rebounds with the recovery of the ridership.

When significant disruptions happen in the system, the point predictions fail miserably, but the uncertainty predictions stay accurate and indicate the changing situation. The MAPE for the test set in Table 2.3 was 18% for CTA and 25% for ridesharing. For the three additional periods, the MAPEs are 93%, 43%, 36% for CTA and 133%, 54%, and 47% for ridesharing, respectively. The uncertainty predictions recovered a lot faster than MAPE. The calibration error returned to pre-COVID levels at the steady recovery stage, although the point prediction error is still 10% higher than before. If only a 95% prediction interval is considered, even at the stay-at-home home stage, we can achieve pretty precise prediction intervals (95.6% and 96.7% for CTA and ridesharing).

The generalizability between different two-parameter distributions is similar, although each has distinct characteristics. TG-GCN excels at low ridership; Lap-GCN is heavy-tailed and more conservative, while HetG-GCN relies on the similarity between training and testing domains. Since the situation is evolving constantly, there is no probabilistic assumption that dominates all scenarios. But some general conclusions can be drawn. In most cases, TG-GCN has the best calibration error due to its left truncation. Enforcing non-negativity is beneficial since the ridership has not recovered to pre-COVID levels for either CTA or ridesharing. Lap-GCN typically produces the best NLL and point prediction due to its more distributed shape, reducing overfitting to pre-COVID data. As ridesharing trips become more spontaneous in post-COVID times, Lap-GCN’s conservative prediction intervals outperform others.

2.5.4 Spatiotemporal Uncertainty

Travel demand uncertainty has both spatial and temporal patterns: spatially, uncertainty is higher at stations with higher volumes, and temporally, uncertainty is higher in the afternoons. Figure 2-6 shows the spatial distribution of predicted uncertainty for CTA Rail at different times of the day during the steady recovery test period (Oct 12 - 25, 2020). Each figure's bottom left corner zooms in on the “loop” in downtown Chicago. Uncertainty is proportional to the size and the color of the circles, with darker and larger circles indicating higher uncertainty.

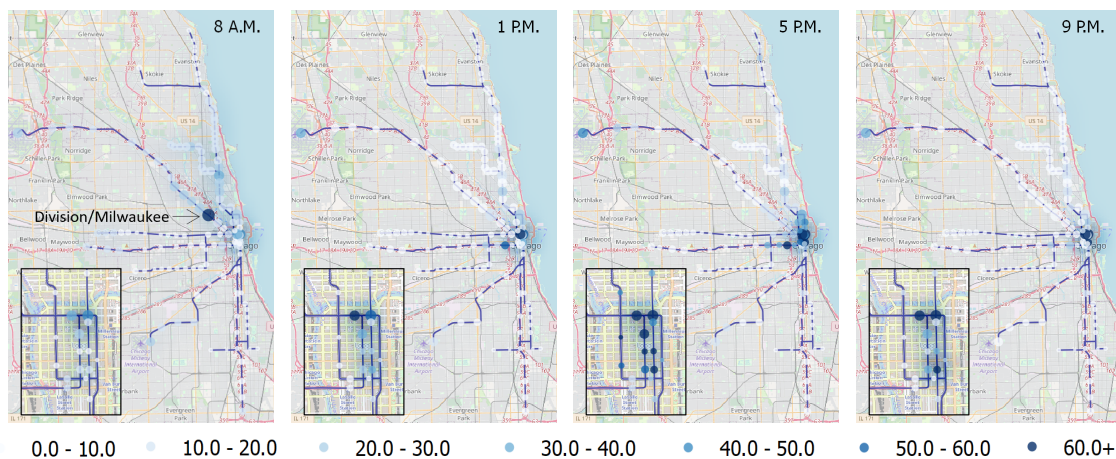


Figure 2-6: Spatiotemporal Uncertainty: Standard deviations of estimated CTA Rail station tap-ins in the 15-minute periods starting at 8 A.M. (morning peak), 1 P.M. (midday), 5 P.M. (afternoon peak), and 9 P.M. (evening).

Spatially, uncertainty is higher at busier stations. During the same time period, darker and larger circles appeared near downtown, transfer stations, and airports. Statistically, if the occurrence of each potential trip is a random Bernoulli process with probability p , and we do not have further information on the values of p for each trip, having more potential trips n will yield higher uncertainty in the sum. The number of observed trips has a binomial distribution with variance $np(1 - p)$, which is proportional to n . Practically, the trips from downtown, transfer stations, and airports are usually more diverse and complex, promoting spontaneity in trip-making resulting in higher uncertainty.

Temporally, uncertainty is higher during the afternoon peak. The uncertainty

during the morning peak is generally smaller than that during midday and evening, although the morning peak has significantly higher ridership. Statistically, this observation could be attributed to having knowledge about some of the p 's. Since the morning peak primarily consists of commuting trips, the probability of the trips happening is higher than recreational trips, which tend to happen during midday and afternoons.

Spatiotemporal uncertainty predictions can inform strategic decision-making regarding capacity buffers. First, the framework could be used to identify bottlenecks and outliers in the system. For example, the station Division/Milwaukee on the blue line has unusually high uncertainty during the morning peak. Further investigation can be done to identify the reasons for the abnormal behavior and perform demand management. Moreover, different strategies are needed for different types of systems. In systems with a fixed, relatively large capacity, such as stations along the same subway line, uncertainty at the busiest stations at peak times is the most critical, as the rest of the line will have quite a lot of excess capacity. Therefore, understanding uncertainty in low-demand regions is important for behavioral analysis but less critical for service planning. However, in systems that are built to meet demand, such as ride-hailing, uncertainty in lower-demand regions is as important as higher-demand regions, as the supply in those regions will be proportionally low. Re-balancing actions will be needed across the system.

2.6 Conclusion

Despite the importance of uncertainty in travel demand prediction, past studies use deep learning to predict only the average travel demand but not quantify its uncertainty. To address this gap, this study proposes a framework of Prob-GNNs to quantify the spatiotemporal uncertainty of travel demand. The framework is concretized by six probabilistic assumptions (HomoG, Pois, HetG, TG, GEns, Lap) and two deterministic ones (GCN and GAT), which are applied to transit and ridesharing data, yielding the following conclusions.

First, the Prob-GNN framework can successfully quantify spatiotemporal uncertainty in travel demand with empirical evidence from the transit and ridesharing datasets. In both cases, the Prob-GNNs can accurately characterize the probabilistic distributions of travel demand while retaining the point predictions of a similar quality to the deterministic counterpart (HomoG). Second, the probabilistic assumptions have a much more substantial impact on model performance than the deterministic ones. The two deterministic architectures lead to less than a 3% difference in NLL loss, while a wise choice of probabilistic assumptions could drastically improve model performance. Specifically, the two-parameter distributions (e.g., heteroskedastic Gaussian) achieve the highest predictive performance, which is 20% higher in log-likelihood and 3-5 times lower in calibration errors compared to the one-parameter baseline distributions. Third, the Prob-GNNs enhance model generalizability under significant system disruptions. By applying the models trained on pre-COVID data to three post-COVID periods, we show that the point predictions fail to generalize, but the uncertainty predictions remain accurate and successfully reflect the evolving situations. Even under significant domain shifts, the difference in predictive performance among the two-parameter distributions is minor. Lastly, Prob-GNNs can reveal spatiotemporal uncertainty patterns. Uncertainty is spatially concentrated on the stations with higher travel volume and temporally concentrated on the afternoon peak hours. In addition, the Prob-GNNs can identify the stations with abnormally large uncertainty, informing real-time traffic controls to address potential system disruptions proactively.

Future studies could advance this work by making further theoretical or empirical efforts. Theoretically, Prob-GNN is a parametric uncertainty quantification method, which should be compared to Bayesian and non-parametric methods regarding prediction quality. Empirically, as transportation patterns are diverse across modes and cities, testing our framework under varying data scales and contexts is important to further corroborate the above conclusions and to better inform policy-making. In addition, our framework can be generally applied to predicting origin-destination flows, travel safety, or even climate risks. Since uncertainty in urban systems has

broad policy implications, future studies could integrate the Prob-GNNs with robust optimization methods to enhance urban resilience.

Chapter 3

Data Fusion: Deep hybrid models to combine numerical data and satellite imagery for travel behavior analysis

3.1 Background

Demand modeling has been a theoretically rich field widely applied to various travel behavioral analyses. Researchers created the multinomial logit model to capture random utility maximization as a decision mechanism [101], the nested logit model to represent the tree structure of the alternatives [102], the mixed logit model to capture the behavioral heterogeneity in preference parameters [103], and the hybrid demand model to reveal the latent behavioral structure [148, 11, 146]. These demand models have been applied to analyze car ownership, travel mode choice, adoption of electric vehicles, and destination choice, among many other travel behaviors [140, 54, 13]. However, the existing demand models use only low-dimensional numeric data, including sociodemographic characteristics and trip attributes, while lacking the capacity to process high-dimensional unstructured data, such as urban imagery. Urban imagery has been shown to contain valuable information on the built environment, socioeconomic factors, and mobility patterns by recent deep learning research [67, 7, 183].

It seems a natural effort to extend the classical demand models to incorporate urban imagery, thus reflecting a more realistic behavioral mechanism and enriching the demand modeling tools.

To operationalize this idea, the key question is *how to integrate the numeric and urban imagery data, leveraging the computational power of deep learning for urban imagery while retaining economic information for practical uses*. On the one hand, demand models have demonstrated that travel decisions depend on travel time, travel cost, income, age, and other numeric data, which facilitates rigorous microeconomic analysis [138, 101]. The microeconomic analysis is valuable because it leverages the random utility theory to compute critical economic parameters such as social welfare and substitution patterns of alternatives [138, 184]. However, an exclusive focus on numeric data misses the tremendous opportunities in the recent big data revolution, in which unstructured data, such as urban imagery, accounts for more than 80% of data growth [8]. On the other hand, the deep learning models in the field of urban computing have used urban imagery - typically satellite or street-view images - to predict sociodemographic characteristics, achieving high predictive performance [67, 42]. However, urban computing research exclusively focuses on using urban imagery for prediction, which dismisses the practical needs of computing elasticity, social welfare, market shares, and other important economic factors. A pure deep learning approach ignores the sociodemographic and travel-related attributes, but it is implausible that travel decisions do not depend on income, age, and travel costs. Since each of the two research paradigms only partially captures behavioral realism, this dichotomy necessitates an effort to integrate them, thus successfully incorporating urban imagery into decision analysis while retaining certain economic information for practical use.

This study presents a synergetic framework of the deep hybrid model (DHM), which is visually represented by a crossing structure with a vertical and a horizontal axis, as shown in Figure 3-1. The vertical axis represents the components in classical demand modeling, including the numeric inputs and outputs (e.g., sociodemographics, travel attributes, and behavioral outputs). The horizontal axis represents deep learning components, specifically an autoencoder that encodes and regenerates urban

imagery. These two axes are connected through a latent space, which serves as the core to integrate the numeric data and urban imagery. This framework is named “deep hybrid” because it resembles and enriches the classical hybrid demand models [11]. It resembles the classical hybrid model because it is similar to the visual diagram of Figure 3 in [11], with our horizontal axis resembling the measurement model and our vertical axis resembling the structural model. It enriches the classical hybrid demand models with deep architectures to construct a latent space with higher dimensions capable of processing unstructured data, such as urban imagery. This framework builds upon the recent efforts in deep choice analysis [153, 157, 155], which illustrates that deep learning enables researchers to extract economic information for practical uses. However, a particular challenge is designing an effective operator to mix the numeric data and urban imagery, thus rendering this latent space predictive and informative.

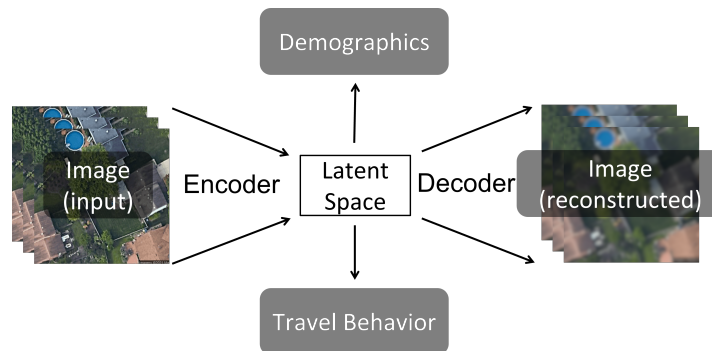


Figure 3-1: Diagram of a deep hybrid model (Model 4-6 in Table 3.1)

This challenge is addressed by designing a mixing operator, which encodes the numeric and imagery data into a latent space with simple concatenation and supervised autoencoders. The latent variables are then imported into a simple behavioral predictor, which is similar to classical demand models. Section 3.2 reviews the literature about travel demand modeling, computer vision (CV), and urban computing applications. Section 3.3 presents the design of the DHM framework with the mixing operators. Section 3.4 introduces data collection, local context, and experiment design. Section 3.5 investigates three empirical questions (1) whether the DHM framework can outperform demand models and deep learning models by effectively integrating the sociodemographics and satellite imagery (Section 3.5.1); (2)

whether the latent space in DHM is spatially and socially meaningful (Section 3.5.2), and (3) DHM can generate new satellite imagery and derive the corresponding economic information for practical uses (Section 3.5.3). Section 3.6 summarizes our findings, limitations, and broad implications. Our scripts have been uploaded to https://github.com/sunnyqywang/demand_image/tree/DHM-AE to promote open science.

3.2 Literature review

3.2.1 Travel demand modelling

Demand models have been used extensively to analyze human decisions regarding travel mode choices [61, 157, 191], adoption of new technologies [44, 159], willingness to pay [9], and behavioral loyalty [133, 66]. The most common demand models are the discrete choice models (DCM) based on the theory of random utility maximization. The data inputs of DCMs include individual- and alternative-specific attributes (e.g., sociodemographics, travel time, and cost) or psychometric features (e.g. perceptions and attitudes), which are often modeled using latent variables. Although the latent variables cannot be directly measured, they can be estimated with structural equation models [146]. To integrate the latent variables and random utility theory, researchers developed hybrid demand models to jointly analyze the observed choice and latent variables, serving as a state-of-the-art demand model [148, 11]. The hybrid demand models have no theoretical limitations in the dimensions of the latent space; however, due to practical constraints such as model estimation algorithms, computational power, and the cost of data collection, the majority of the existing work adopted less than five latent variables without application to urban imagery [139, 56, 97, 29, 62].

Increasingly, researchers started to enhance the demand models by accounting for complex input-output relationships with deep learning [20, 88, 58, 143, 169, 170, 193, 194]. Deep neural networks (DNNs) take advantage of the increasingly available

computation power and use gradient descent and regularization techniques (such as L1, L2, dropout, early stopping, and more) to enable the training of large networks with flexible architecture design. As a result, DNNs have shown superior performance in various applications [71, 191]. However, DNNs are often criticized for the interpretability and instability problems [114, 57], which could be mitigated by gradient-based methods [157, 142, 4] or integrated DCM-DNN architectures [167, 154, 137]. The integrated architectures can facilitate researchers to derive economic information, enhance training efficiency, and stabilize the learning process. For example, the DNNs can fit the residuals of the DCM backbone [167, 154, 137], learn decision rules [141], and generate choice sets [176]. DNNs can also learn latent representations from survey indicators [166], personal characteristics and their interactions with the alternative attributes [51], GPS trajectories [35, 178, 177], and embeddings for categorical variables that are typically hard to handle in traditional frameworks [5]. However, all existing studies use only numeric data, while deep learning is powerful because it can learn from high-dimensional data. There were some attempts at leveraging high-dimensional data, including treating GPS trajectories as images [35, 178], and stacking multiple features along a route to form 2D inputs [177], but none have leveraged real urban imagery. Real urban imagery is associated with various urban and socioeconomic characteristics, as shown by the vast number of studies in computer vision and urban computing.

3.2.2 Computer vision and urban computing

Rich urban and socioeconomic characteristics are contained in various urban images, including nightlight, satellite, and street view images. Using satellite imagery, researchers can learn land-use patterns [2], quantify green cover [134], measure physical appearances of neighborhoods [108], and predict socioeconomic status [67, 7, 183, 42]. In transportation, researchers used urban imagery to monitor traffic flows during the pandemic [26], predict pedestrian intentions using car and traffic cameras [124], and predict the usage of active modes based on street-view images [52]. Several studies also used street-view images to learn the perceived safety and general attitudes

towards neighborhoods [109, 33, 185]. Reviews of urban imagery applications can be found in [65, 16]. However, we have not identified any study that used observed human decisions such as travel behavior, as modeling outputs.

Besides the predictive models, researchers are paying increasing attention to the generative models. As a baseline, the autoencoder (AE) can learn latent representations from images using an encoder and generate new images using a decoder. Even in its simplest form, the AE can learn latent representations, which are effective for prediction tasks [125, 172]. A baseline AE can be enhanced by the multitask supervised learning as a regularization in its latent space, thus further improving its capacity of representation learning [32, 63]. However, AE is criticized for lacking sampling capacity and generating blurry images, so variational autoencoders (VAE) and generative adversarial networks (GAN) were proposed as two remedies. The VAE regularizes the latent space of AE with a Gaussian prior, enabling a smooth transition of generated images [50]. The GAN can generate realistic images through the game-theoretical training of its discriminator and generator. The loss terms in GAN can be augmented to the AEs to improve the manifold learning [98, 34, 173] and the quality of image generation [84, 14, 113]. The state-of-the-art generative models integrate the loss functions of VAEs and GANs into the baseline AE, marking the convergence of the two lines of research [37, 128]. Despite the proliferation of generative models in CV literature, no study has investigated how to generate satellite images from sociodemographic and travel characteristics.

More important than using travel behavior as another application, we must design a novel approach drastically different from the existing urban computing research to model human decisions. Human decision is more complex than built environment labels (e.g. trees, parking lots, or other land use patterns) because it inevitably involves the discussions of social heterogeneity, economic implications, and human decision mechanisms. Unlike sociodemographic data, individual pixels in urban imagery do not have socioeconomic meanings, posing a unique challenge for deriving socioeconomic information from urban imagery. To address these challenges, we propose the DHM framework that encodes urban imagery and sociodemographics into a latent

space, by which we could conduct associative analysis between sociodemographics, travel decisions, and satellite imagery.

3.3 Theory

3.3.1 General framework of deep hybrid models

The DHM can be represented as:

$$y_n = g(z_n) = g(\mathcal{M}(x_n, I_n)) \quad (3.1)$$

in which x_n and I_n represent the numeric and imagery inputs, and y_n represents the travel behavioral outputs. The two key components in DHM are $\mathcal{M}(x_n, I_n)$, which combines the numeric sociodemographics x_n and the urban imagery I_n , and $g(z_n)$, which predicts the travel outputs. In other words, the DHM framework consists of a *mixing operator* $\mathcal{M}(\cdot)$ and a *behavioral predictor* $g(\cdot)$. The behavioral predictor follows a generalized linear form: $g(z_n) = \sigma(\beta' z_n)$, in which $\sigma(\cdot)$ represents the link function and $\beta' z_n$ is a linear transformation of the latent variables $z_n = \mathcal{M}(x_n, I_n)$. In this formulation, $g(\cdot)$ is significantly simplified so that we can concentrate the discussion on the mixing operator. Meanwhile, $g(\cdot)$ is also flexible enough to accommodate a variety of output categories: single outputs, soft choice probabilities, and discrete choices.

Table 3.1 summarizes six models for the mixing operator $\mathcal{M}(x_n, I_n)$. We will provide an overview of the table in this section, followed by a detailed discussion of mixing operators in Section 3.3.2. In Table 3.1, the first column explains the formula for computing the latent variables $z_n^{(i)}$. The second column presents the optimization formulation of neural network parameters. $E_\phi(\cdot)$ represents the encoder network parameterized by ϕ , $D_\psi(\cdot)$ represents the decoder network parameterized by ψ , and $F_\omega(\cdot)$ represents the sociodemographic supervision network parameterized by ω . The third column presents the latent dimensions of $z_n^{(i)}$, which vary drastically across the six models.

Table 3.1: Design of the mixing operator in deep hybrid models

Mixing operator	Optimization for parameter estimation	Latent dim
Panel 1: Benchmark Models		
Model 1: Only sociodemographics (SD)		
$z_n^{(1)} = x_n$	N/A	10
Model 2: Only imagery with autoencoders (AE)		
$z_n^{(2)} = E_{\hat{\phi}}(I_n)$	$\hat{\phi}, \hat{\psi} = \underset{\phi, \psi}{\operatorname{argmin}} \mathcal{L}_{rec}$	18,432
Panel 2: Deep Hybrid Models		
Model 3: Combining Models 1-2		
$z_n^{(3)} = [z_n^{(1)}, z_n^{(2)}]$	N/A	10 + 18,432
Model 4: Supervised autoencoder (SAE)		
$z_n^{(4)} = E_{\hat{\phi}}(I_n)$	$\hat{\phi}, \hat{\psi}, \hat{\omega} = \underset{\phi, \psi, \omega}{\operatorname{argmin}} \lambda \mathcal{L}_{rec} + (1 - \lambda) \mathcal{L}_{sup}$	18,432
Model 5: Supervised autoencoder with enhanced image regeneration (latent dim = 10)		
$z_n^{(5)} = E_{\hat{\phi}}(I_n)$	$\hat{\phi}, \hat{\psi}, \hat{\omega} = \underset{\phi, \psi, \omega}{\operatorname{argmin}} \lambda (\alpha_{rec} \mathcal{L}_{rec} + \alpha_{lips} \mathcal{L}_{lips} + \alpha_{GAN_G} \mathcal{L}_{GAN_G} + \alpha_{KL} \mathcal{L}_{KL}) + (1 - \lambda) \mathcal{L}_{sup}$	10
Model 6: Supervised autoencoder with enhanced image regeneration (latent dim = 4,096)		
$z_n^{(6)} = E_{\hat{\phi}}(I_n)$	Same as Model 5	4,096

Models 1 and 2 are benchmark models. Model 1 incorporates only sociodemographics x_n without any imagery. Model 2 incorporates only the imagery information without including any sociodemographics by using the neurons of a baseline autoencoder as the latent variables. Autoencoders learn latent representations z_n from image I_n using an encoder $z_n = E_{\hat{\phi}}(I_n)$, and reconstruct the image \hat{I}_n using a decoder $\hat{I}_n = D_{\hat{\psi}}(z_n)$, with ϕ and ψ representing the parameters in the encoder and decoder. The basic autoencoder in Model 2 is trained with the reconstruction loss:

$$\mathcal{L}_{rec} = \sum_n |I_n - \hat{I}_n| \tag{3.2}$$

, which measures the L1 distance between the input and reconstructed images in the pixel space. Although the reconstruction loss is limited to only the pixels, other loss terms can be added to regularize the latent variables and improve the quality of representation learning (See Models 4-6).

3.3.2 Mixing operator: supervised autoencoders

Model 3 represents the simplest form of DHMs because it linearly concatenates the latent variables of Models 1 and 2. Although concatenation is simple, it is widely used to integrate diverse data sources in deep learning literature [82, 105]. By comparing the performances of Model 3 to Models 1-2, we could observe whether the two data sources are complementary. This simple concatenation in Model 3 is a benchmark of all the DHMs.

Models 4-6 are designed as the supervised autoencoders (SAE), the diagram of which is visualized in Figure 3-2. Different from a baseline AE, the latent variables z_n of SAEs can reconstruct images through the decoder $D_\psi(z_n)$ and sociodemographics through a supervision network $F_\omega(z_n)$. The supervision of sociodemographics enhances the stability in the latent space because it extends the basic AE to multitask learning with the sociodemographic supervision loss as regularization [86]. Both multitask learning and regularization can constrain the latent space and improve training stability [24, 100].

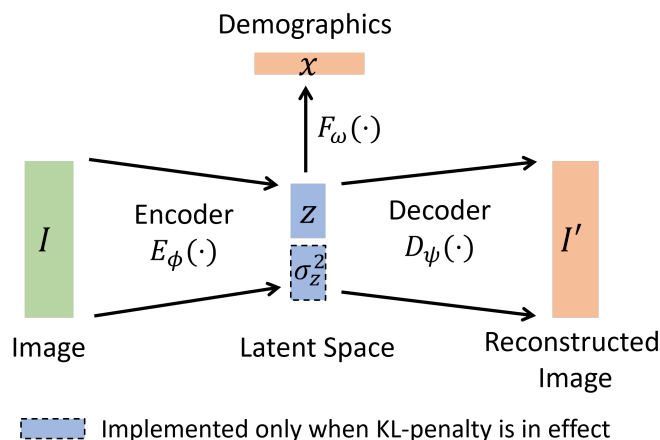


Figure 3-2: Supervised Autoencoders

The SAEs (Models 4-6) are trained by minimizing the sum of an AE loss \mathcal{L}_{AE} and a sociodemographic supervision loss \mathcal{L}_{sup} .

$$\mathcal{L}_{SAE} = (1 - \lambda)\mathcal{L}_{AE} + \lambda\mathcal{L}_{sup} \quad (3.3)$$

in which \mathcal{L}_{sup} represents the absolute error of the sociodemographic variables x weighted by an adaptive hyperparameter α_{sup} :

$$\mathcal{L}_{sup} = \alpha_{sup}|x - \hat{x}| \quad (3.4)$$

In Equation 3.3, a mixing hyperparameter $\lambda \in [0, 1]$ measures the trade-off between image reconstruction and sociodemographic supervision. When $\lambda = 0$, the latent space contains only imagery information. When $\lambda = 1$, the latent space contains only sociodemographic information. Equation 3.3 represents the most general form of the loss function shared across Models 4-6, but the AE loss \mathcal{L}_{AE} varies between Model 4 and Models 5-6.

Model 4 is a baseline SAE using the L1 reconstruction loss (Equation 3.2) as the \mathcal{L}_{AE} . Therefore, the loss function of Model 4 is:

$$\mathcal{L}_{SAE} = (1 - \lambda)\mathcal{L}_{rec} + \lambda\mathcal{L}_{sup} \quad (3.5)$$

Although Models 5-6 share the same general diagram (Figure 3-2) and training objective (Equation 3.3) as Model 4, they significantly expand upon the AE loss \mathcal{L}_{AE} into a list of loss terms, which can effectively enhance the quality of image generation. The AE loss in Models 5-6 includes the reconstruction loss \mathcal{L}_{rec} , the KL divergence loss between the estimated latent space distribution and the standard normal prior distribution \mathcal{L}_{KL} , the learned perceptual image patch similarity (LPIPS) \mathcal{L}_{lpiPs} , and the adversarial objective \mathcal{L}_{GAN_G} and \mathcal{L}_{GAN_D} [128]. The KL-penalty is introduced in the same way as VAE to enable smooth transitions and sampling in the latent space and constrain the magnitude of the variance in the latent space. The LPIPS loss and the adversarial objective enhance the realism of the generated images and mitigate the blurriness caused by relying solely on pixel-space metrics such as the L1 reconstruction loss. The full breakdown is shown in Equation 3.6:

$$\mathcal{L}_{AE} = \alpha_{rec}\mathcal{L}_{rec} + \alpha_{KL}\mathcal{L}_{KL} + \alpha_{lpiPs}\mathcal{L}_{lpiPs} + \alpha_{GAN_G}\mathcal{L}_{GAN_G} \quad (3.6)$$

in which all the α terms are the adaptive hyperparameters, which will be introduced later. Both Models 5 and 6 use Equation 3.6 as the objective function, although they are designed with different latent dimensions (10 vs. 4,096) to test the impacts of latent dimensions on prediction and image generation. In Equation 3.6, The first loss term is still the reconstruction loss \mathcal{L}_{rec} as Equation 3.2 - the pixel level L1 distance between input and reconstructed images.

The KL divergence \mathcal{L}_{KL} measures the distance between the calculated and the standard Gaussian distributions, as shown in Equation 3.7. The latent variables use k -dimensional diagonal Gaussian distribution with estimated mean \hat{z}_n and variance $\hat{\sigma}_{z_n}^2$.

$$\mathcal{L}_{KL} = \sum_n KL(\mathcal{N}(\hat{z}_n, \hat{\sigma}_{z_n}^2) || \mathcal{N}(0, 1)) = \frac{1}{2} \sum_n \sum_k (\hat{z}_{n,k}^2 + \hat{\sigma}_{z_{n,k}}^2 - 1 - \log(\hat{\sigma}_{z_{n,k}}^2)) \quad (3.7)$$

The \mathcal{L}_{lips} loss - learned perceptual image patch similarity (LPIPS), also known as perceptual loss \mathcal{L}_{lips} - is widely used to enhance perceptual similarity between the deep features of two images [187]. To calculate the LPIPS loss between the original and reconstructed images I_n and \hat{I}_n , the feature stacks from L layers are extracted from a pre-trained neural network and unit-normalized in the channel dimension, which is designated as $\hat{y}_{I_n, hw}^l, \hat{y}_{\hat{I}_n, hw}^l \in \mathbf{R}^{H_l \times W_l \times C_l}$. The features are then scaled by vector w_l before computing the L2 distance. Its formula is:

$$\mathcal{L}_{lips} = \sum_n \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}_{I_n, hw}^l - \hat{y}_{\hat{I}_n, hw}^l)\|_2^2 \quad (3.8)$$

Lastly, the generative adversarial loss uses a discriminator network $Disc(\cdot)$ to differentiate between real and generated images. Equation 3.9 shows the generator (decoder D_ψ) loss function \mathcal{L}_{GAN_G} , and Equation 3.10 shows the discriminator loss function \mathcal{L}_{GAN_D} . GANs are trained in a game-theoretic manner with alternating generator and discriminator updates. At each step, the model first trains the generator (decoder), the encoder, and the supervision network while fixing the discriminator

network weights. It then trains the discriminator with the other models fixed.

$$\mathcal{L}_{GAN_G} = \sum_n -Disc(\hat{I}_n) \quad (3.9)$$

$$\mathcal{L}_{GAN_D} = \sum_n \max(0, 1 + Disc(I_n)) + \max(0, 1 - Disc(\hat{I}_n)) \quad (3.10)$$

The loss terms are associated with adaptive hyperparameters, including α_{sup} , α_{rec} , α_{KL} , α_{lips} , and α_{GAN_G} . These adaptive hyperparameters dynamically balance the loss terms and stabilize the complex training process. For example, α_{sup} is calculated by the ratio of the loss terms' gradients with respect to the parameter magnitude in the last layer L of the decoder D_ψ [37] using $\delta = 10^{-6}$ for numerical stability.

$$\alpha_{sup} = \frac{\nabla_{G_L}[\mathcal{L}_{rec}]}{\nabla_{G_L}[\mathcal{L}_{sup}] + \delta} \quad (3.11)$$

The α_{GAN_G} is also adaptive and calculated similarly as α_{sup} .

$$\alpha_{GAN_G} = \frac{\nabla_{G_L}[\mathcal{L}_{rec}]}{\nabla_{G_L}[\mathcal{L}_{GAN_G}] + \delta} \quad (3.12)$$

To simplify the training process, other adaptive hyperparameters are fixed using the default setting from the stable diffusion models¹, where $\alpha_{rec} = 1$, $\alpha_{KL} = 1e^{-6}$, and $\alpha_{lips} = 1$. The adaptive hyperparameters α 's and the static hyperparameter λ serve different purposes: the former mainly stabilizes the training process, and the latter diagnoses the trade-off between image reconstruction and sociodemographic supervision.

Models 4-6 use the SAE design to interact with sociodemographics and satellite imagery through the latent space. Although the formula $E_{\hat{\phi}}(I_n)$ contains only images as inputs, the encoder parameters $\hat{\phi}$ are trained using both x_n and I_n . Therefore, the $\hat{\phi}$ absorbs the information from two data structures, so it enables the latent variables $E_{\hat{\phi}}(I_n)$ to blend information from sociodemographics and urban imagery. The design of the SAE models is highlighted due to its similarity to the classical hybrid

¹https://github.com/CompVis/stable-diffusion/blob/main/configs/autoencoder/autoencoder_kl_8x8x64.yaml

demand model structure [11]. In classical hybrid demand models, a structural component describes how the sociodemographic variables relate to travel behavior, and a measurement model describes the relationship between observed variables and latent variables [148, 11]. Our DHMs resemble this structural approach mainly through the multitask learning framework, and we further enrich the existing hybrid models by leveraging the higher-dimensional deep architecture to design the latent space and incorporate urban imagery.

3.3.3 Behavioral predictor

The behavioral predictor $g(z)$ is designed as a generalized linear regression:

$$\hat{y} = g(z) = \sigma(\beta'z) \quad (3.13)$$

Despite its simplicity, Equation 3.13 is sufficiently flexible to accommodate three output variables: (1) aggregate travel mode shares as individual outputs, (2) aggregate travel mode shares as a joint output, and (3) individual travel mode choices. A general form of training the behavioral predictor is:

$$\operatorname{argmin}_{\beta} \mathcal{L}(y_n, \sigma(\beta'z_n)) + \theta \|\beta\|_p \quad (3.14)$$

in which $\mathcal{L}(y_n, \sigma(\beta'z_n))$ represents the loss function (e.g. mean squared error or negative log-likelihood), β represents the parameters, θ is the sparsity hyperparameter to control the sparsity of β . In all three examples, the sparsity hyperparameter θ is the weight for L1 (LASSO) regularization to address overfitting and identify the relevant latent variables. The LASSO approach is essential in reducing the potential overfitting from the high-dimensional latent space in Models 2, 3, 4, and 6 in Table 3.1. The three outputs are clarified in the following three subsections.

Aggregate travel mode shares as separate outputs

The first example is relatively straightforward: since the outcome y_n is a continuous scalar to represent the aggregate travel mode shares, the link function σ is designed

as an identity mapping:

$$\sigma(\beta' z_n) = \beta' z_n \quad (3.15)$$

The training uses the mean squared error as the objective.

Aggregate travel mode shares as a joint output

The second example uses the aggregate travel mode shares as a joint output of the model. Since the joint mode shares should add up to one, a softmax function is specified as the link function to implement this constraint. Using the linear transformation on $\beta' z_n$, we could represent the output mode shares as:

$$P_{nk} = \frac{e^{V_{nk}}}{\sum_j e^{V_{nj}}} = \frac{e^{\beta'_k z_n}}{\sum_j e^{\beta'_j z_n}} \quad (3.16)$$

in which P_{nk} represents the market shares of mode k in region n . Since the market shares follow a probability distribution, Kullback-Leibler (KL) loss is used to specify the general loss in Equation 3.14:

$$\mathcal{L}_{KL} = \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K P_{nk} (\ln P_{nk} - \ln \hat{P}_{nk}) \quad (3.17)$$

Disaggregate travel mode choice

The third example uses individual travel mode choice as the model outputs. The choice probabilities of trip n for alternative k are specified as

$$P_{nk} = \frac{e^{V_{nk}}}{\sum_j e^{V_{nj}}} \quad (3.18)$$

where P_{nk} is the probability of trip n taken with alternative k , V_{nk} is the utility of alternative k for trip n . Since the disaggregate travel mode choice involves origin-destination pairs, the alternatives' attributes for each OD pair concatenate the origin o_n and destination d_n : $z_n = [z_{o_n}, z_{d_n}]$. Similar to the aggregate travel mode analysis, the utility function takes a simple linear form as $V_{nk} = \beta'_k z_n$. However, since discrete choice models have unique data structures in the input variables, which consist

of individual- and alternative-specific variables, the detailed specification is slightly different from the aggregate travel mode analysis (Appendix A.1). Let y_{nk} represent the observed mode ($y_{nk} = 1$ if mode k is used for trip n , and $y_{nk} = 0$ otherwise), and N is the total number of trips. The training loss in Equation 3.14 is substantiated by the cross entropy loss:

$$\mathcal{L}_{ce} = \frac{1}{N} \sum_{t=1}^N \sum_{k=1}^K -y_{nk} \ln P_{nk} \quad (3.19)$$

3.3.4 Deriving economic information from generated satellite imagery

Recent work has demonstrated that deep learning models can provide economic information as complete as the classical discrete choice models [157]. Building upon deep learning, the DHMs can empower researchers to compute economic information from generated satellite images. Here, we provide the formula for computing market shares, substitution patterns of alternatives, social welfare, and choice probability derivatives with highlights on the latent variables z that connect images, sociodemographics, and travel behaviors. Specifically, the market share of alternative k can be computed as

$$s_k = \sum_n P_{nk} = \sum_n \frac{e^{V_{nk}}}{\sum_j e^{V_{nj}}} \quad (3.20)$$

Social welfare of individual n takes the standard log sum formula:

$$\frac{1}{\alpha_n} \log\left(\sum_j e^{V_{nj}}\right) \quad (3.21)$$

where α_n measures the marginal utility of income that translates social welfare into dollar values. The substitution pattern for two alternatives j and k is defined as the ratio of their choice probabilities:

$$\frac{P_{nj}}{P_{nk}} = \frac{e^{V_{nj}} / \sum_{k'} e^{V_{nk'}}}{e^{V_{nk}} / \sum_{k'} e^{V_{nk'}}} = e^{V_{nj} - V_{nk}} \quad (3.22)$$

Although the three equations above appear similar to those in the standard demand models, they already incorporate satellite imagery to compute the utility value V_{nk} through the latent variable z . Among the six models in Table 3.1, the DHMs (Models 3-6) embed both satellite images and sociodemographic variables into the latent variable z_n . Although only images are used as inputs in Models 4-6, the parameters $\hat{\phi}$ are the functions of both sociodemographics and images, thus encoding their information into the latent variables $E_{\hat{\phi}}(I_n)$. Any latent variable \tilde{z} can be used to visualize satellite imagery through the decoder as $\tilde{I} = D_{\psi}(\tilde{z})$ and compute the utility, market shares, social welfare, and substitution patterns with Equations 3.20, 3.21, and 3.22.

Using the DHMs, we could further compute a directional gradient of choice probabilities regarding the latent variable z :

$$\nabla_u P_{nk}(z) = u \cdot \nabla P_{nk}(z) \quad (3.23)$$

where $\nabla P_{nk}(z)$ represents the gradient of choice probability regarding the latent variable z , and u represents a direction to move in the latent space. Since DHMs connect z and urban imagery I , this directional gradient can describe the sensitivity of choice probabilities with respect to the satellite image I_n corresponding to the latent variable z_n . For example, the direction can be defined as one-directional as $u = z_2 - z_1 = E(I_2) - E(I_1)$ by using two existing images I_1 and I_2 , or extended to a multi-directional movement by linearly combining multiple u 's: $u = a_1 u_1 + a_2 u_2$, where u takes into account two directions u_1 and u_2 . Our empirical analysis will demonstrate the directional sensitivity of choice probabilities regarding satellite images by discretizing the latent space. For example, we can create a directional vector $u = z_2 - z_1$ to define the directional movement and compute the choice probabilities and satellite images using a new latent variable $\tilde{z} = z_1 + au = z_1 + a(z_2 - z_1)$, in which a ranges between zero and one. With this approach, we can visualize a new satellite image \tilde{I} through the decoder as $\tilde{I} = D_{\psi}(\tilde{z})$ and compute the choice probabilities through the behavioral predictor as $g(\tilde{z})$. We can also compute market shares,

social welfare, and substitution patterns by applying $V_{nk}(\tilde{z})$ to Equations 3.20, 3.21, and 3.22.

3.4 Experiment design

3.4.1 Data

The experiments combine satellite images from Google API, socio-demographics from the census, and individual travel behavior from MyDailyTravel Chicago Survey in 2018-2019 [1]. The same number of satellite images were sampled from each census tract, which is then linearly combined in the latent space of the DHM for each region s by $z_s = \frac{1}{I_s} \sum_{i=1}^{I_s} z_i$, where z_s is the latent representation averaged over I_s satellite images. The sociodemographics are obtained from the American Community Survey (ACS)², which includes total population, age groups, racial composition, education, economic status, and travel information (e.g. commuting time). After data preprocessing, The dataset has 1,571 census tracts with 80K observed trips. Figure 3-3 shows five samples of the standard satellite images with a 256×256 size.

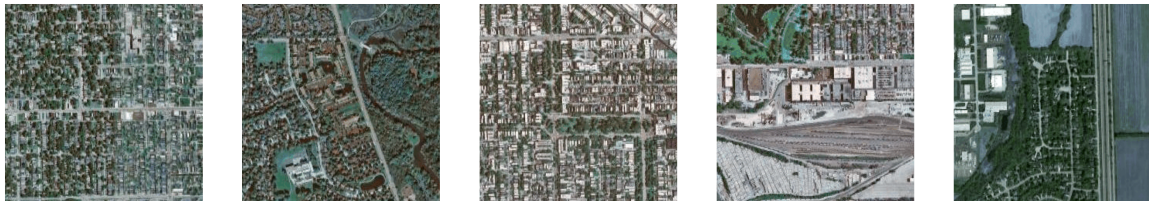


Figure 3-3: Samples of satellite images used in the experiment.

The MyDailyTravel survey provides individual sociodemographics and trip attributes for the disaggregate analysis. It contains 42 variables, including time of the trip (morning/after), trip distance, home-based trips, trip purposes, perceived time importance, age, disability, and education, among many other travel and social factors. The initial travel modes have more than ten categories, but they are aggregated into auto, active (walk+bike), public transit, and others, thus creating a relatively balanced choice set.

²<https://www.census.gov/programs-surveys/acs/data.html>

3.4.2 Model training

The training of DHMs takes two steps - training the mixing operator and then the behavioral predictor - as summarized in Figure 3-4. During stage I, we train the SAEs, which consists of the encoder $E_\phi(I)$, decoder $D_\psi(z)$, and the sociodemographic predictor $F_\omega(z)$. Every input image is encoded into a latent variable z , which then passes through $F_\omega(z)$ and $D_\psi(z)$ to reconstruct sociodemographics and images. During stage II, only the behavioral predictor $g(z)$ is trained. In testing, an input image is encoded into a latent variable z , which then predicts sociodemographics and travel behaviors through $F_\omega(z)$ and $g(z)$, and generates images through $D_\psi(z)$.

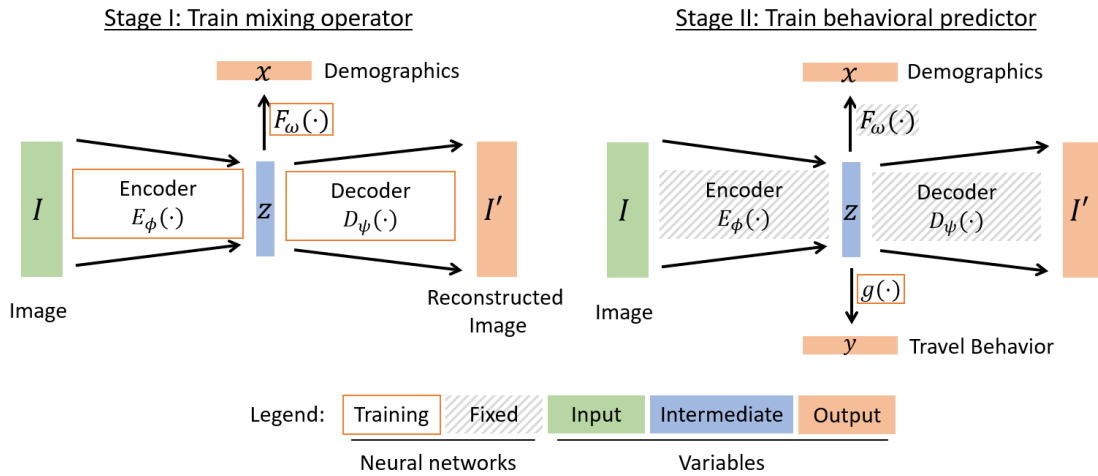


Figure 3-4: Two-step training

The mixing hyperparameter λ is searched on a linear scale ($\lambda = 0.1, 0.3, 0.5, 0.7, 0.9$) to evaluate the impacts of the mixing on model performance. The sparsity hyperparameter θ is searched through all positive values to evaluate the sparsity effects. In model evaluation, the five-fold cross-validation is used. The R^2 is calculated separately on the auto, active, and public transit to evaluate the aggregate models. The cross-entropy loss and accuracy rates are used to evaluate the disaggregate models.

The six models require drastically different levels of computational resources. Models 2-4 are trained on a NVIDIA 2080 Ti GPU with 11GB RAM for the 52 million parameters in the ResNet-50 architecture [53]. The Models 5-6 are trained on a NVIDIA V100 GPU with 32GB RAM, which costs around \$15,000 at the current

market price, for the 126 million parameters in the U-Net from the stable diffusion models. Besides monetary costs, the relatively simple model (e.g. Model 4) takes around 12 hours to train, and the more complex ones (e.g. Model 6) take two to three days. Such a significant amount of computational resources could inhibit the academic communities from replicating this study. Therefore, we present all six models with distinctly different degrees of complexity, thus facilitating the replication of at least Models 1-4 for the researchers with limited computational resources. In fact, the advanced Model 6 does not guarantee a better result on all the fronts. As shown below, it dominates others by the quality of image regeneration but generates predictive performance nearly equivalent to that of Model 4.

3.5 Results

3.5.1 Predictive performance

Table 3.2 summarizes the empirical results of the six models in three travel behavior tasks. Panels 1-2 present the aggregate travel demand analysis using travel modes as separate and joint outputs, and Panel 3 the disaggregate travel demand analysis. The aggregate models (Panels 1 and 2) typically use ten sociodemographic variables and 18,432- or 4,096-dimensional latent variables to represent urban imagery. The disaggregate analysis (Panel 3) incorporates an additional 42 travel attributes into the input variables. The six columns correspond to the six models in Table 3.1. The first row in each panel illustrates the dimension of latent variables. Each entry reports the training/testing performance, selected by the highest testing performance with the optimal λ and θ values using five-fold cross-validation. The hyperparameter selection with other λ and θ values can be found in Appendix A.2. Because of its superior quality in image regeneration, only the Model 6 with $\lambda = 0.7$ is used for the further analysis in Sections 3.5.2 and 3.5.3. The predictive performance yields five major findings as follows.

First, both the sociodemographics and satellite imagery are informative for travel

Table 3.2: Predictive performance

Model	1 SD	2 AE	3 [SD AE]	4 Basic SAE	5 SAE (10)	6 SAE (4096)
<i>Panel 1: Aggregate Mode Choice as Separate Outputs - Linear Regression</i>						
Latent Dim	10	18432	10+18432	18432	10	4096
Auto (R^2)	0.553/0.541	0.627/0.532	0.655/0.594	0.720/0.644	0.579/0.566	0.676/0.640
Active (R^2)	0.456/0.446	0.542/0.407	0.463/0.472	0.583/0.531	0.472/0.457	0.566/0.524
PT (R^2)	0.443/0.424	0.503/0.395	0.543/0.464	0.545/0.482	0.462/0.443	0.520/0.483
<i>Panel 2: Aggregate Mode Choice as a Joint Output - Multinomial Regression</i>						
Latent Dim	10	18432	10+18432	18432	10	4096
KL Loss	0.146/0.149	0.149/0.158	0.135/0.146	0.111/0.128	0.142/0.144	0.120/0.129
Auto (R^2)	0.544/0.535	0.514/0.484	0.596/0.557	0.704/0.643	0.577/0.568	0.669/0.637
Active (R^2)	0.448/0.439	0.430/0.396	0.500/0.452	0.627/0.524	0.471/0.453	0.566/0.515
PT (R^2)	0.424/0.407	0.396/0.364	0.476/0.428	0.569/0.476	0.436/0.427	0.543/0.496
<i>Panel 3: Disaggregate Mode Choice - Discrete Choice Analysis</i>						
Latent Dim	42+10*2	18432*2	42+10*2+18432*2	42+18432*2	42+10*2	42+4096*2
CE Loss	0.423/0.422	0.659/0.652	0.380/0.389	0.378/0.409	0.405/0.415	0.373/0.404
Accuracy	0.856/0.857	0.756/0.759	0.871/0.868	0.868/0.855	0.862/0.859	0.871/0.856
Note: each entry is represented as training/testing performance.						

behavioral analysis because the predictive performance of Models 1-2 is substantially higher than zero across all three panels. In Panel 1, Model 2 using only satellite imagery can explain 53.2%, 40.7%, and 39.5% of the variations in automobiles, active travel modes, and public transit (Column 2), slightly lower than 54.1%, 44.6%, and 42.4% in Model 1. This finding still holds in Panel 2, while Model 2 performs much worse than Model 1 in the disaggregate analysis (Panel 3). The results are quite reasonable: urban imagery is slightly less informative than sociodemographics in predicting aggregate travel behavior, while it is much less so than sociodemographics and particularly travel attributes, which are the major incentives determining individual travel behavior. Regardless of these variations, the performance of Models 1-2 demonstrates that both sociodemographics and satellite imagery can assist in travel behavioral prediction.

Second, the numeric data and satellite imagery are complementary, as shown by the performance comparison in Table 3.2 and the information decomposition in Figure 3-5. Across all three panels, Models 3-6 consistently outperform Models 1-2, indicating that it is more effective by mixing rather than independently using the two

data structures. Since Model 3 simply concatenates the latent variables of Models 1-2, its higher performance is attributed to data complementarity rather than our DHM design, which exists only in Models 4-6. Besides comparing performance, the non-zero coefficients in Model 3 also indicate data complementarity. Figure 3-5 illustrates how the non-zero coefficients decrease with a larger sparsity hyperparameter θ . With the optimal $\theta = 2e^{-4}$, the non-zero coefficients originate from both the sociodemographic and urban imagery sides, consisting of six sociodemographic variables and 335 imagery variables, demonstrating the complementary contributions to prediction from the two data structures.

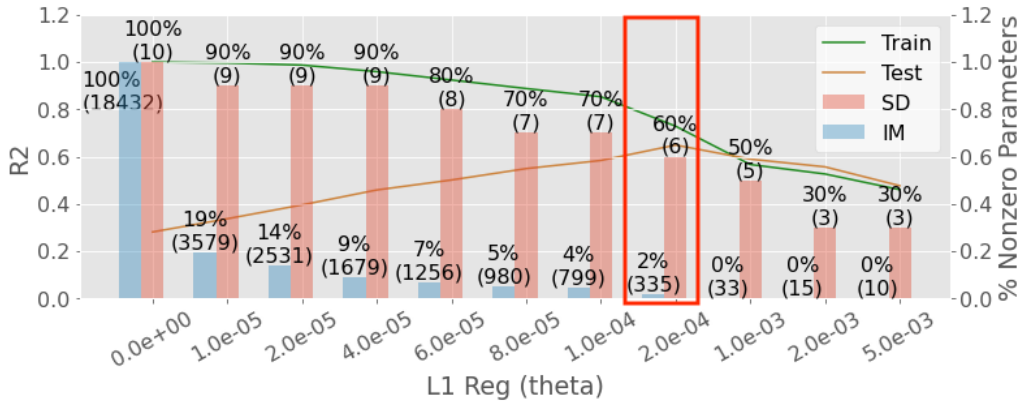


Figure 3-5: Non-zero coefficients from sociodemographics and urban imagery in Model 3 with various θ values.

Third, the SAEs are more effective than the simple concatenation in the aggregate contexts (Panels 1-2) but less so in the disaggregate contexts. Model 4 outperforms Model 3 by 5-10% in R^2 and about 12% in the KL loss in Panels 1-2. This result suggests that the nonlinear mixing of sociodemographics and urban imagery through the SAEs is more effective than a linear concatenation in the aggregate analysis. Across the three travel modes, the performance improvement is more significant in the automobile and active modes, suggesting that the built environment represented by satellite imagery is more informative for using private vehicles, walking, and cycling than public transit. The effects of SAEs appear less evident in the disaggregate analysis because the linear concatenation in Model 3 outperforms Models 4-6, suggesting that the travel attributes influence individual travel choices in a linear way.

Fourth, it is important to create a relatively high-dimensional latent space (Model 6) for high predictive performance, although the low-dimensional latent space (Model 5) can also explain substantial variation in travel behaviors. Models 5-6 were designed to be identical in data, model, and training but differed in only the latent dimensions (10 vs. 4,096). Across the three panels, Model 6 consistently outperforms Model 5 in predictive performance, suggesting the importance of a relatively high-dimensional latent space for effective prediction. However, Model 5 also achieves substantial predictive performance, reaching at least 80-90% of the predictive performance in Model 6. This finding suggests that the explanatory power in the latent variables has a somewhat long-tail distribution: it is concentrated in the first few principal dimensions but also spread into hundreds of others.

Lastly, the DHMs can achieve relatively high performance for predicting the sociodemographic variables when the λ value gradually deviates from zero. Figure 3-6 visualizes how the R^2 in predicting sociodemographics varies with λ for every sociodemographic variable. Since λ balances the image reconstruction and the sociodemographic prediction, a higher λ value can instill more sociodemographic information into the latent space. As shown in Figure 3-6, a relatively high λ value can significantly improve the predictive performance of the sociodemographics, including population density, age, racial shares, and income. With an optimal λ value at around 0.5-0.7, the predictive performance achieves a relatively high R^2 (around 80%) in population density, racial structures, education, and income, but relatively lower R^2 (around 30-50%) in age and average travel time.

3.5.2 Navigating the latent space

With the design of DHMs, the latent space encodes the sociodemographic and urban imagery information. The latent space can be understood by using dimension reduction techniques, including K-Means to identify the discrete latent clusters and t-distributed stochastic neighbor embedding (tSNE) for continuous embedding.

We identify five clusters with distinct spatial, sociodemographic, and land use patterns using K-Means. Figure 3-7 shows the image samples from each cluster. The

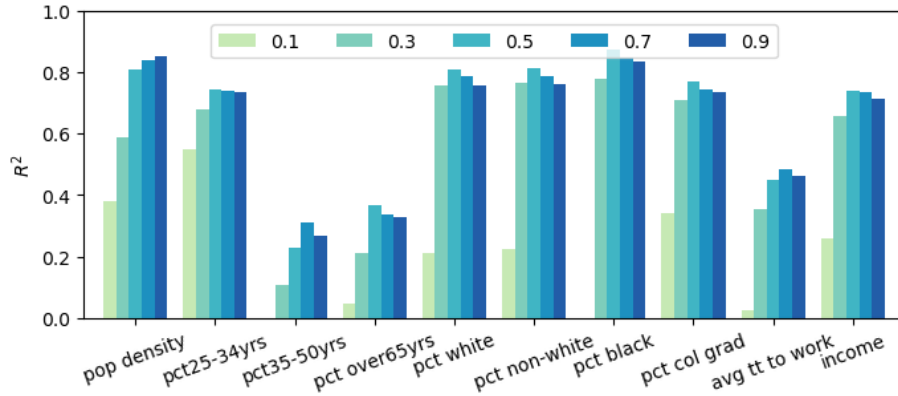


Figure 3-6: Performance of predicting sociodemographics in Model 6 with various λ values

first cluster represents a suburban town with substantive residential areas cut through by a major highway. The second cluster represents the high-density neighborhoods in the downtown area. The third cluster represents a suburban town center with a more industrial presence. The fourth mixes the high-density urban region with some public facilities of large footprints. The fifth cluster is a typical suburban region with very low density and large green space.

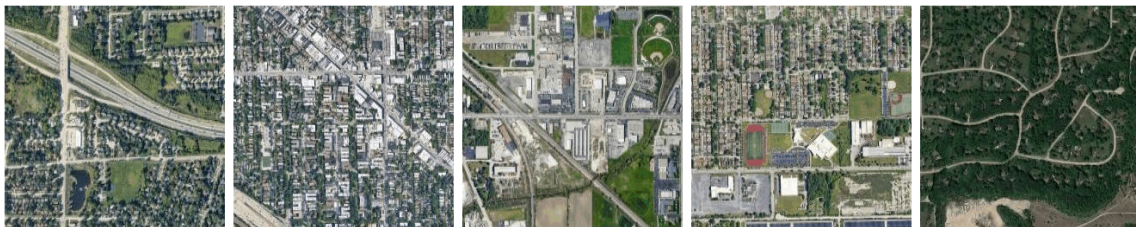


Figure 3-7: Image samples of cluster centers

Although spatial information is not explicitly used as inputs, the five clusters can reveal spatial clusters, such as Chicago’s urban vs. suburban regions, as shown in Figure 3-8. The first cluster in northwest Chicago represents the suburban residential regions. The second cluster identifies downtown Chicago. The third and fourth clusters represent southern Chicago, with the former being closer to downtown. The fifth cluster represents the outskirts surrounding Cook County.

Using tSNE, the latent space could be simplified as a 2D visualization, which can demonstrate its association with the continuous sociodemographic variables. The five

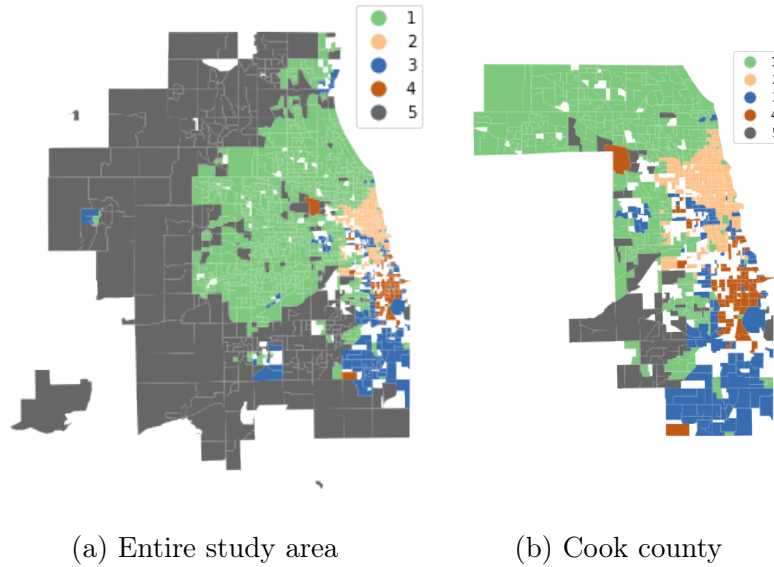


Figure 3-8: Spatial distribution of clusters

clusters are marked in the latent space in Figure 3-9a, and the continuous sociodemographics in Figure 3-9b-f. Such continuous patterns in sociodemographics provide additional information to the discrete clusters. For example, the sociodemographics of the first cluster (north) correspond to the high-income and high-education people. The second cluster (downtown) has high income, high population density, a high percentage of adults and college graduates, and moderate to low commuting time. The sociodemographics of the third and fourth (south) clusters correspond to low income, low population density, low proportions of college graduates, and long commuting times. The last cluster (outskirts) has the lowest density without much sociodemographic richness.

3.5.3 Deriving economic information for generated satellite imagery

The latent space in DHMs enables us to generate new satellite images and derive economic information from them. The process will be demonstrated by starting with three existing images in Figure 3-10, which represent the census tracts in the South, North, and Central Chicago areas, the sociodemographics of which are shown in Table

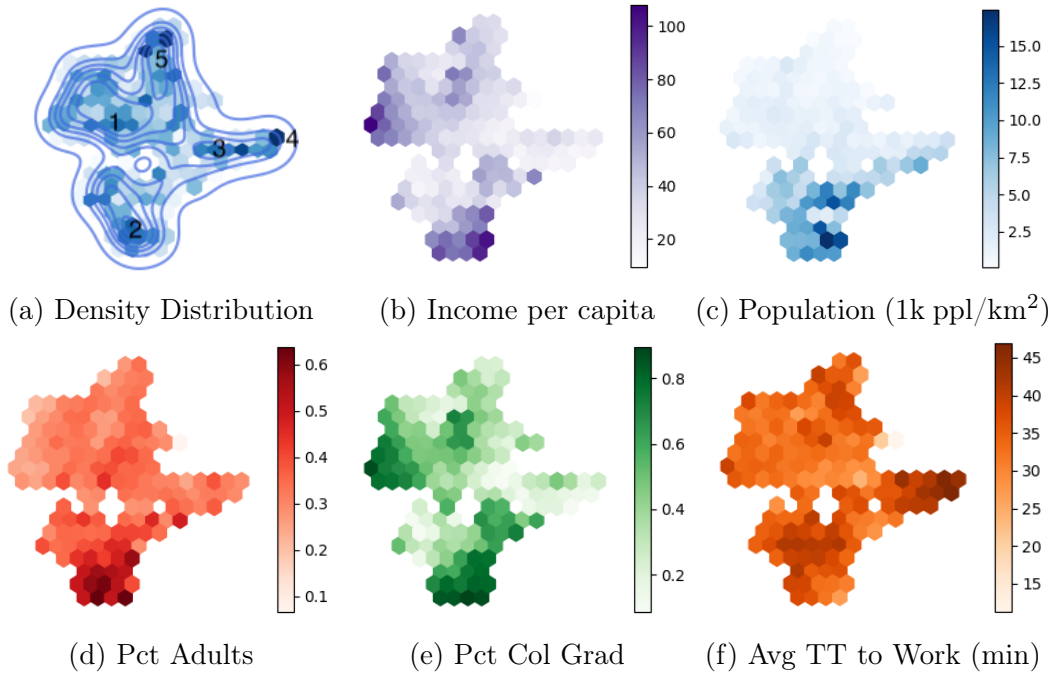


Figure 3-9: tSNE visualization of latent space

3.3. The first image presents a mixed urban texture with a major highway in southern Chicago. This area is associated with low income, low education, and long commute times. The other two are located in the north suburban and central Chicago areas. The north region represents the high-income and low-density areas, while the central region is denser and relatively wealthy. Northern Chicago is represented by a typical suburban pattern with curved roads and low-density buildings, while central Chicago by small-scale and grid-shaped building blocks. The latent variables of the three census tracts are denoted as z_s , z_n , and z_c , and their corresponding images as I_s , I_n , and I_c .

Table 3.3: Sociodemographics of three representative census tracts

Census Tract	Pop Density (k-ppl/ km^2)	% College Grad	Avg Time to Work (min)	Income (10k/capita)	Auto	PT	Active
South	2.5	10.7%	33.6	12.1	69.9%	15.4%	5.17%
North	1.0	78.2%	37.1	90.8	86.6%	5.1%	6.5%
Central	6.2	75.9%	28.0	70.8	12.0%	46.7%	38.4%



Figure 3-10: Satellite images of three representative census tracts

Deriving economic information with one-directional image generation

New satellite images can be generated using latent variable \tilde{z} through the decoder $\tilde{I} = D_\psi(\tilde{z})$. Specifically, the latent variable \tilde{z} is created by adding a directional vector u to the source latent variable z_s : $\tilde{z} = z_s + a_1 u$, in which a_1 ranging from zero to one in 0.2 increments, and u is the direction between the target and source image $u = z_c - z_s$ in the latent space. As a result, a new satellite image \tilde{I} can be generated with every \tilde{z} as shown in Figure 3-11.

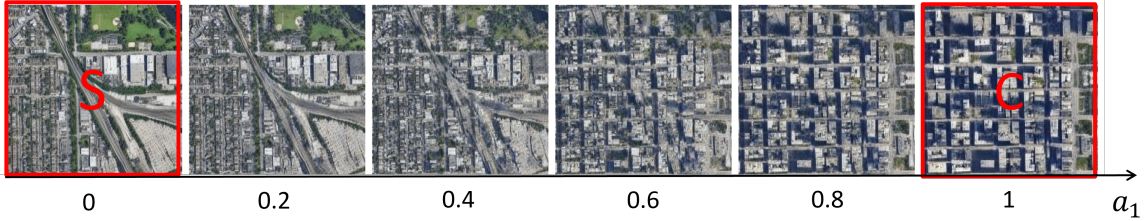


Figure 3-11: One-directional image generation

For a generated satellite image, its corresponding socioeconomic information can be derived through the DHM framework, including travel behavior $g(\tilde{z})$, utility values $V_{nk}(\tilde{z})$, and sociodemographics $F_\omega(\tilde{z})$. Figure 3-12 visualizes how market shares, social welfare, substitution patterns, probability derivatives, and sociodemographics vary with the movement in the latent space indicated by the scale factor a_1 . The market share of auto mode, computed as $e^{\beta'_k \tilde{z}} / \sum_j e^{\beta'_j \tilde{z}}$, decreases when the latent variable moves from the central to the south Chicago. The social welfare, computed with the logsum form $\frac{1}{\alpha_n} \log(\sum_j e^{\beta'_j \tilde{z}})$, significantly decreases when the utilities of travel modes are roughly equivalent. The derivatives of choice probabilities $\nabla_u P_{nk}$ are marginally decreasing when a specific travel mode dominates the mobility market. Overall, larger

a_1 values are associated with more white people and college graduates, higher income levels, and lower travel time to work.

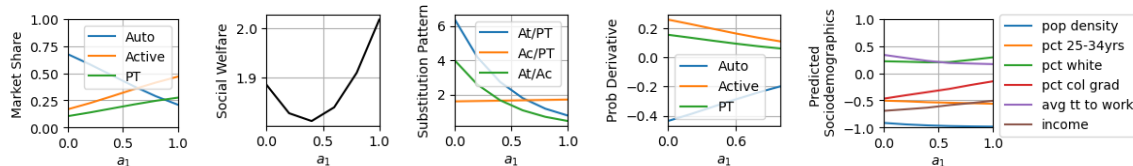


Figure 3-12: Economic information of images generated along one direction (Southern to Central Chicago: $\tilde{z} = z_s + a_1(z_c - z_s)$)

Deriving economic information with two-directional image generation

The one-directional image generation can be extended to a two-directional one using $\tilde{z} = a_1u_1 + a_2u_2 = a_1(z_c - z_s) + a_2(z_n - z_s)$. The two directions are defined by $z_c - z_s$ and $z_n - z_s$, representing the movement from the south to north and central Chicago in the latent space. Figure 3-13 visualizes a 6×6 matrix of the generated images, in which only three images (S, N, and C) exist in reality (Figure 3-10) while the other 33 images are generated. This two-directional image generation approach generally applies to other satellite images. Another example is shown in Appendix A.3.

The generated urban images have high visual quality, as shown in Figure 3-13 and 3-14. Moving from the south to central Chicago, the grid structure and dense buildings gradually dominate the new urban landscape. When $a_1 = 0.6$ and $a_2 = 0.0$ (Figure 3-14a), the generated image has the uniform urban grid, resembling central Chicago, with mixed green areas and corridors permeating into this urban area. Moving from the south to north Chicago, the major north-south green belt is gradually replaced by curved suburban roads and low-density buildings. When $a_1 = 0.0$ and $a_2 = 0.6$ (Figure 3-14b), the urban landscape retains the curved suburban road networks similar to north Chicago, but with mixed building footprints.

Similar to the one-directional examples, the socioeconomic information can be derived for every generated satellite image, as shown in Figure 3-15. The satellite image on row i and column j in Figure 3-13 corresponds to the market share and social welfare values in the pixel on row i and column j in Figure 3-15. Such matrix



Figure 3-13: Two-directional image generation

plots in Figure 3-15 can provide a holistic view of the two-directional interactions. For example, public transit usage does not change significantly when moving from southern to northern Chicago but varies when moving from central Chicago. Beyond market shares and social welfare, it is also feasible to compute other economic parameters for the generated image \tilde{I} . Here, we skip further details and leave them for future studies.

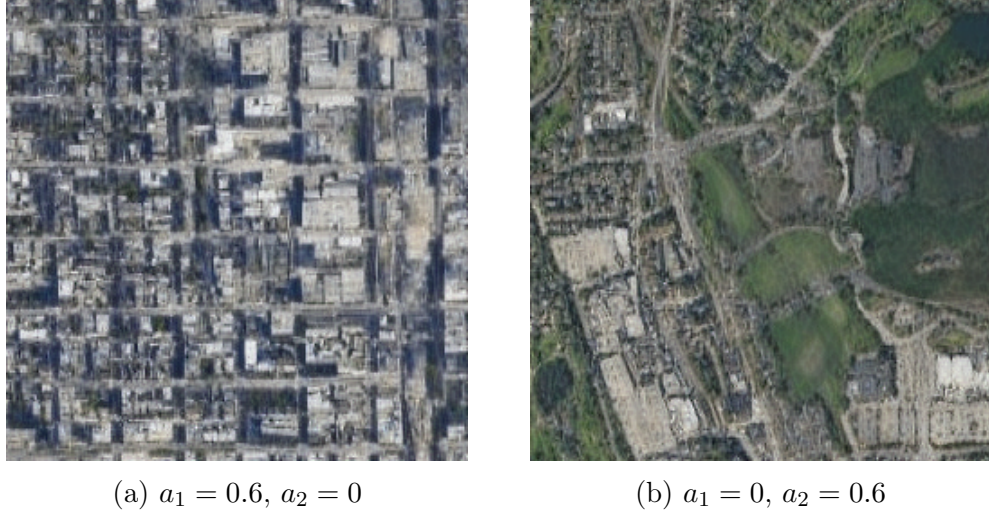


Figure 3-14: Two generated satellite images

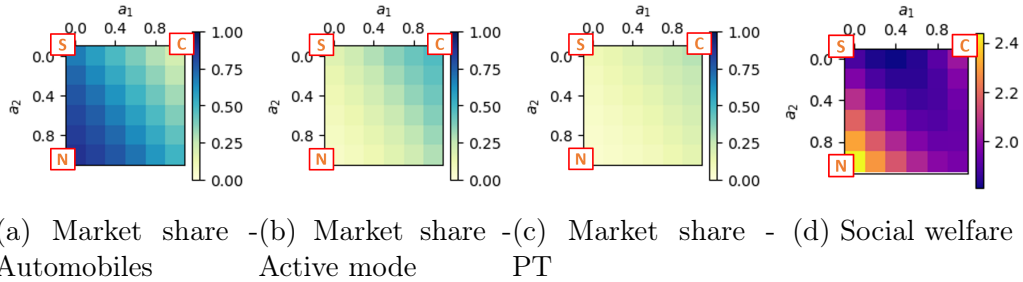


Figure 3-15: Economic information of images generated along two directions

3.6 Conclusion

Travel decisions can be influenced by the factors represented by either numeric data or urban imagery. Although classical demand modeling cannot effectively process urban imagery, expanding on this capacity is imperative due to the explosion of unstructured data. This study proposes the DHM framework to integrate numeric and satellite imagery by combining the analytical capacity of the demand modeling and the computational capacity of the DNNs. After empirically examining its performance, we demonstrate three major findings of DHMs, which correspond to three contributions of this research.

First, this study demonstrates the complementarity of data and model. The data complementarity is demonstrated by the higher performance of the DHMs combining numeric and imagery data over the benchmark models. The model complemen-

tarity is demonstrated by integrating classical demand modeling and deep learning through the latent space in the DHMs. Technically, the complementarity is achieved by the SAE design. Unlike the simple concatenation, SAEs combine the imagery and sociodemographics through the supervised sociodemographics for AE. Second, this study enriches the hybrid demand models with deep hybrid models by constructing a latent space from the satellite imagery using deep architectures. This latent space contains meaningful social and spatial characteristics, although its interpretation is no longer as straightforward as the classical hybrid demand models. Third, DHMs offer one method to compute economic information for the generated satellite imagery. The high-quality images are generated using the linearly transformed latent variables through the AE in stable diffusion models. The economic information related to the generated urban images (e.g. mode shares and social welfare) can be computed because the DHMs successfully associate sociodemographics, travel behavioral outputs, and satellite imagery through the latent space.

Our framework is hybrid for five reasons. Three reasons are already fully elaborated: it is hybrid because it integrates numeric and imagery data, combines classical demand modeling and deep learning, and enriches the classical hybrid model family. Two other reasons, relatively less elaborated so far, are on the machine learning side. The DHMs are hybrid because they mix supervised and unsupervised learning: the vertical axis of DHMs in Figure 3-1 represents supervised learning, while the horizontal autoencoder represents unsupervised learning. It also mixes the discriminative and generative methods. The decoder generates new satellite imagery, while the behavioral prediction is discriminative. The two perspectives are explained in other machine learning studies that adopted the hybrid generative-discriminative or supervised-unsupervised methods [85, 46, 31, 90].

The DHM framework still has many limitations. The DHMs leverage the unique capacity of deep autoencoders and stable diffusion models to process urban imagery. However, this strength is inevitably associated with the challenges in a highly complex neural network. Although deep learning can minimize errors effectively, the final solution by no means is globally optimal. Neural networks converge to various

local minima, and this non-identification issue is still a pending challenge in deep learning research. The economic information computed from the generated satellite imagery is a numerical approximation, as opposed to the analytical solutions from the classical parameter-based discrete choice models. The latent space in DHMs is no longer constrained to a few latent variables, limiting its interpretability in the classical parameter-based sense. It is also ambiguous how to guarantee stability and robustness in such a high-dimensional latent space. The DHMs achieve a relatively high predictive performance, but their transferability could be limited. It is unclear whether the models can still perform well when trained in one context and used in another. This study uses the autoencoder in the state-of-the-art stable diffusion for image generation, but the quality of image generation can always be further improved. Meanwhile, the U-Net architecture in Models 5-6 is computationally expensive, thus limiting its replicability for future researchers. Regarding future work, the DHMs combine two purposes: representation learning for prediction and image regeneration for image-based story-telling. But, it might not be easy to improve two purposes simultaneously, so future studies can focus on achieving one purpose without sacrificing the other. Empirically, researchers could explore how to use satellite imagery in other contexts. For example, sociodemographics are often considered exogenous variables for energy consumption, health, and air pollution, so future research could combine satellite imagery with sociodemographics to analyze these factors using the DHM framework.

Chapter 4

Generative Urban Design: Human-guided automatic urban design via diffusion models

4.1 Background

Urban design, a term originating from the Latin word “urbs” (meaning “city”), has gained significant attention as cities grow and evolve. Although the precise philosophical definition of urban design remains debatable, its core objective is widely accepted as “shaping better places for people” [23]. Urban design is crucial in achieving public welfare, sustainable development, efficient use of resources, and long-term economic growth. Therefore, urban design is highly interdisciplinary, drawing on fields such as social science, civil engineering, environmental engineering, economics, architecture, and political science. In practice, urban design is highly complex and iterative. Although the process may vary depending on the specifics of each project, there are a few stages in general. First, the objectives and goals for the project must be identified. Then, a land survey is performed to assess the existing infrastructure, natural environment, and the sociodemographics of the neighborhood. The urban planners will identify strategies for achieving the goals based on the assessment. The strategies

will need to be communicated with a multitude of stakeholders for feedback and iteration, including the local government, the community residents, the urban planners, and the real estate developers. Effective communication among these stakeholders is paramount, and it is particularly crucial to engage the public in shaping their neighborhoods, as they possess invaluable knowledge about their local contexts and needs [74, 107].

“A picture is worth a thousand words.” Visualizations can facilitate stakeholder communication, providing intuitive views of the plan to the non-experts. Kevin Lynch’s seminal work on the “imageability of the city” underscores the significance of mental maps and visual perception in shaping urban environments [96]. Quick and realistic visualizations are pivotal in bridging gaps in understanding, fostering effective communication, and achieving stakeholder consensus.

Given the iterative nature of urban design, tools that expedite design, communication, and feedback are highly desired. In recent years, Artificial intelligence (AI) technologies have exhibited exciting potential in various aspects of urban design, including gaining data-driven insights, evaluating and optimizing the performance of plans, and creating visualizations of plans. One distinct advantage AI has over humans is the ability to learn from large amounts of data. AI models have been built to produce deep insights, identify trends and patterns in complex city ecosystems, and generate optimal land use and building layouts [149, 150, 190]. The insights gained from these models help human planners make more informed and data-driven decisions. Additionally, AI can help evaluate and optimize the performance of plans. For example, Delve¹ from Sidewalk Labs and Forma² from Autodesk are two commercial tools that are capable of generating and evaluating designs. Lastly, the advent of generative AI brought more potential for AI to learn and apply visual styles for quick visualization of completed plans with image-to-image models [181, 36].

Although AI has huge potential to transform the urban design process, there are still many practical challenges. First, current research has mainly focused on specific

¹<https://www.sidewalklabs.com/products/delve>

²<https://www.autodesk.com/products/forma/overview>

and well-constrained tasks. As urban design comes in different shapes and forms, parameterizing such a process is not straightforward. Although the consensus is that the future will be human-driven and machine-assisted, there is no overarching framework describing how this human-machine collaboration will manifest itself. Second, generative AI consumes large amounts of data, but labeled data in the urban setting is expensive and, hence, relatively scarce. Finally, as this technology is still emerging, AI tools are primarily found in research environments and have limited practical application.

In this work, we make attempts at addressing the challenges by envisioning a human-machine collaboration framework for the urban design process, then instantiating this framework with a model trained from data automatically labeled from existing resources, and lastly evaluating the state-of-practice and reflecting on challenges and future research directions. The rest of the chapter is structured as follows: Section 4.2 reviews the literature on the role of imagery in urban design, the relevant AI technology, and the application of AI in urban design. Section 4.3 proposes a vision for the human-machine collaboration in urban design. Section 4.4 instantiates this framework with a specific use case, followed by the detailed account of methodology in Section 4.5 and dataset in Section 4.6. Section 4.7 provides a quantitative evaluation of the model from a user survey. Section 4.8 conducts a qualitative demonstration of the model, showcasing the model’s generation capabilities with respect to constraints and land use descriptions. Section 4.9 discusses the feedback from practitioners and scholars, and reflects on the limitations and potential extensions. Lastly, Section 4.10 concludes this chapter.

4.2 Literature Review

4.2.1 Urban Imagery in Planning and Design

In the realm of urban planning and design, the significance of urban imagery, or visualizations, has been widely recognized. Urban imagery is critical in the initial

design stages and can improve public understanding and participation to facilitate effective communication and consensus building [96, 10]. Historically, designers have crafted maps, blueprints, and visualizations by hand, a labor-intensive process often relying on artistic interpretation. In the last decade, the scientific community began to harness the power of imagery for predictive purposes. Both satellite imagery and street view imagery were shown to be correlated with various sociodemographic and economic indicators [67, 7, 127, 183, 42]. Although deep learning models effectively improve the understanding of communities, there is a large gap between model development and model deployment for decision-making [19]. Generative AI may revolutionize how we envision and build our urban environments. Its capacity to produce coherent natural language descriptions and vivid urban imagery offers a powerful means of enhancing communication, making complex concepts more accessible. At the same time, Generative AI capitalizes on the advantages of deep learning and thus has the potential to shape a future where urban development is both visionary and data-informed.

4.2.2 Image Generation Models

With the rise of deep learning and neural networks, tremendous progress has been made with image synthesis. The paradigm has shifted multiple times, from variational autoencoders to generative adversarial networks, and now to diffusion models.

As a baseline, the autoencoder (AE) can learn latent representations from images using an encoder and generate new images using a decoder. However, the baseline AE lacks sampling capacity and generates blurry images. Variational autoencoders (VAE) regularize the latent space of autoencoders with a Gaussian prior, enabling a smooth transition of generated images [50]. However, VAEs still have limitations in generating high-quality images.

Generative adversarial networks are known for generating high-quality, realistic images. GAN consists of two networks, a generator and a discriminator. A game-theoretic (adversarial) training scheme is used to update the two networks in alternate steps. The discriminator tries to distinguish between real samples and the generator-

produced samples, whereas the generator tries to produce samples that can hinder the discriminator. Introducing a discriminator and this training scheme to any generator network leads to the generation of highly realistic images [84, 14, 113]. As a result, GANs have shown tremendous success in various types of image synthesis tasks. For example, DCGAN is one of the earliest, and most popular GAN that generates realistic images from a random vector; StyleGAN is known for its ability to control the “styles” of the generated images [72]; CycleGAN is designed specifically for image-to-image generation. Despite the success of GANs, there are still a lot of challenges associated with the adversarial training process [132]. First, GAN training can be highly unstable. Since updates are made to two networks at the same time, the balance between the generator and the discriminator is very delicate and sensitive to architectures and hyperparameters. Second, GAN suffers from mode collapse, where the generated samples lack diversity. As the generator finds the principal mode and focuses around it, the other modes in the data distribution do not get learned properly. A lot of effort has been put into solving these issues. Despite promising achievements, these issues persist in the general setting since they stem from the training paradigm.

At the same time, diffusion models emerged as a more powerful paradigm in image synthesis. Diffusion models are a class of likelihood-based models that generate images by gradually removing noise from a signal [59, 110, 30]. The training procedure involves two processes: the forward diffusion process and the reverse diffusion process. The forward diffusion process is a Markov chain, where noise is added to the training sample incrementally until pure Gaussian noise is achieved. The reverse diffusion process aims to reverse the forward process. A neural network is trained to remove the noise at each step until we recover the original training sample. Therefore, at inference time, images are generated by running the reverse diffusion process on random noise. Compared to GANs, diffusion models exhibit better scalability and parallelization, as well as more stable training and higher fidelity images. The only drawback is that diffusion models take up much more computational resources and time at both training and inference [28]. With rapid improvements in hardware, diffusion models now form the backbone of the state-of-the-art models, such as OpenAI’s DALL-E2

[123], Google’s Imagen [130], and the open-sourced stable diffusion [129].

To date, image synthesis models have been applied to a wide variety of tasks, which can be divided into two categories: text-to-image and image-to-image. Earlier image synthesis models used numerical vectors to prompt the generation, which is difficult to interpret and greatly slowed the adoption of these models in the broader scientific community. OpenAI’s Contrastive Language-Image Pre-training (CLIP) [122] can learn visual concepts from natural language supervision efficiently. The emergence of CLIP bridged the gap between language and images, contributed to cross-modal understanding, and enabled text-guided image synthesis, which lowered the barrier of entry to experiment with the models and popularized generative AI in the broader scientific community. Another important category is image-to-image applications, including image editing [6], image inpainting [95], image-to-image translation [38], image super-resolution [131], etc. Among these models, latent diffusion, also known as stable diffusion, is one open-sourced model that is capable of performing both text-to-image and image-to-image tasks [129]. The latent space richness has inspired researchers to fine-tune and extend stable diffusion to discover more applications and study its properties. The latent space in stable diffusion has also been shown to be useful in segmentation [165], classification [195], and anomaly detection [119]. To introduce capabilities beyond the original stable diffusion, ControlNet has provided a versatile gateway towards fine-tuning with custom conditions beyond text [186]. The ControlNet framework demonstrated the ability to perform in-context learning, with promising generalization capability on controllable image editing [162, 121].

4.2.3 Generative Urban Design

With the rapid advances in generative AI, applications of deep generative models in urban design and design have been widely explored. Deep generative models have demonstrated impressive performance and great potential in various tasks. Compared to humans, AI can learn from vast amounts of data very quickly and has the potential to learn hidden patterns not defined by humans from real design cases. AI can also facilitate stakeholder engagement in the early stages because of fast prototyping and

visualization of extensive design options.

There are two major approaches to urban design with generative models: designing the functional form and designing in the pixel space. The functional form of land use configurations can be formulated as a longitude-latitude-channel tensor, with the channels being different land use types. A pix2pix model was trained to generate land use type, floor-to-area ratio, and building cover ratio from road network sketches using GAN [115]. To enhance the coherence of the generated plans, a spatial graph can be used to learn the representations of surrounding contexts. Then, a generative adversarial network named LUCGAN was trained to generate a land-use configuration tensor automatically for an empty geographical area [150]. This model was further enhanced with spatial hierarchy, sub-area dependency, and human instructions [149]. In addition to GANs, reinforcement learning can also be used to learn the land use configuration tensors [190]. Designing the functional form enables easier computation of metrics such as spatial efficiency, land use diversity, and walkability. Designing in the pixel space attempts to generate the figures of the final design using image synthesis algorithms. With powerful image generation algorithms, many studies focused on visual exploration have emerged. For example, image-to-image generative networks are trained to predict building footprint from land cover [3]. Additionally, given the street view segmentation, researchers developed tools to generate real-time rendering of satellite images [181, 36] and street view images for users to interact with [112]. Designing in the pixel space makes the design easier and more intuitive in communication while sacrificing some precise functional form information.

Comprehensive reviews of current applications can be found in [64, 168, 68]. It is commonly believed that humans and machines have complementary strengths, and the future of generative urban design will be human-led and machine-assisted, with a strong emphasis on human-machine collaboration [64, 68]. As this field is still in its early stages, many challenges remain. First, current research has mainly focused on specific and well-constrained tasks. There is no overarching framework describing how this human-machine collaboration will manifest itself. Second, as urban design comes in different shapes and forms, parameterizing a design is not straightforward.

Third, generative AI consumes large amounts of data, but labeled data in the urban setting is expensive and, hence relatively scarce.

4.3 The Human-machine Collaboration Vision

To address the challenges identified in the previous sections, we designed an end-to-end framework to facilitate both the design and the visualization process. This process is inspired by recent advances of GenAI in both natural language processing and image synthesis and the multi-modal functionalities of generative models. Figure 4-1 illustrates the human-machine collaboration planning process. Under this framework, the AI takes multi-modal inputs describing the design context. Aided by powerful language models, humans can now describe the problem for the AI to digest. The humans will translate various design factors such as zoning, environment, and demographics into text descriptions. Information that is not easily communicated in text form, such as the built and natural environment can be supplied in image form. The AI takes in both types of information and generates designs for the human planner to evaluate. The designs can be a single output, such as satellite imagery rendering, or a suite of outcomes, including street views, land use blocks, etc. The human planner can then adjust the model inputs and iterate until the desired output is produced.

This framework aims to draw out the best of both human planners and generative models, where humans work on problem definition and decision-making, and AI focuses on rapid prototyping of design ideas. The ultimate goal is to communicate with this AI in the same way as ChatGPT. Imagine a stakeholder meeting scenario where planners and developers meet with the community. The planner starts with a piece of land and takes the geographical and policy constraints to the AI. The AI generates an initial design for feedback. The conversation then begins. The kids and the seniors might want a recreational playground; the commuters might ask for a transit station; the environmentalists might want to make the buildings greener; the developers will want to build to increase the profit on cost, etc. The AI then takes in

the conversation, along with the initial geographical and policy constraints, to iterate on the designs and present alternatives showcasing the tradeoff between different components.

The most critical challenge in the way of turning this vision into reality is the availability of good data. GenAI can only be generated if trained with enough relevant, high-quality samples. In the next section, we present a specific use case of this framework, detailing the data curation, model specification, and the training process.

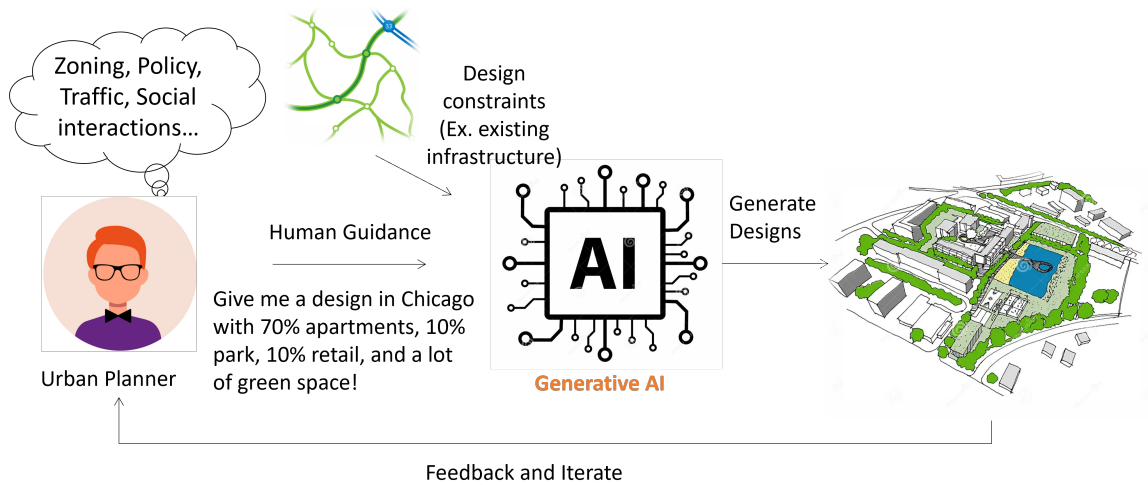


Figure 4-1: The human-machine collaboration planning process.

4.4 Problem Framing

This section instantiates the framework with a specific use case, providing a detailed account of the training process for a single application. In this thesis, a generative design tool is trained for early-stage concept planning.

At this stage, planners conduct preliminary assessments to define goals and establish concept plans. Concept plans are preliminary high-level visual representations that offer a comprehensive perspective on potential development options. We want to foresee how different areas could be allocated for residential, commercial, or industrial purposes, and how much amenities and open spaces to leave space for to meet communal and recreational needs. Simultaneously, we consider rough outlines of

transportation networks, envisioning how residents will navigate the urban landscape. Concept plans often serve as a starting point for public engagement and discussions and lay the ground for detailed design work. Therefore, satellite imagery is chosen as this application’s output due to its ability to offer an intuitive and broad-scale visual representation.

Several target functionalities have been identified to achieve the goals.

1. **Having control over AI-generated land use patterns.**

This states the core function of the model. Since the primary objective is to explore various land use scenarios, humans direct the AI on the specific land use mix to be examined. The major land use types considered are residential, commercial, industrial, park, nature reserve, open parking, and farmlands. Residential areas consist of houses and apartments. Commercial areas include offices, shops, retail, malls, etc. Industrial areas include warehouses and factories. Parks are urban green spaces, mostly within the city. Nature reserves are large natural habitats such as national parks and reserves. Open parking is parking lots without cover, which is visible from aerial imagery. Farmland is, by its name, agricultural.

2. **Producing alternate designs for the same inputs.**

There is a trade-off between more precise control and more AI creativity. On one end of the spectrum, we specify the function and footprints of each building and ask AI for a visualization. This approach is suitable at the final rendering phase, once all detailed designs have been finished. Conversely, we give AI absolute freedom to generate images without any guidance. This approach is suitable for conceptualizing fictional cities. Considering our model’s intended role to provide inspiration in the concept planning phase of real-world projects, we only control for the types of land use patterns but leave the spatial placements to the AI model.

3. **Respecting existing infrastructure and natural environment.**

Elements of the built or natural environment that are not subject to redesign must be accounted for. The model should acknowledge and adhere to these existing constraints while generating new designs.

4. Learn and apply local urban textures in different cities.

One advantage AI has over humans is its ability to learn hidden patterns and styles from data directly. Despite the distinct style inherent to each city, articulating and applying these unique styles can be challenging. We want to leverage AI’s advantage to envision alternate realities for our cities and learn from the best practices.

4.5 Methodology

This section introduces a two-stage workflow designed to fulfill the objectives presented in Section 4.4, as depicted in Figure 4-2. The first stage involves the design of human-guided controls, and the second stage involves selecting and training a model that responds to these controls. A detailed discussion of both stages follows in the remainder of this section.

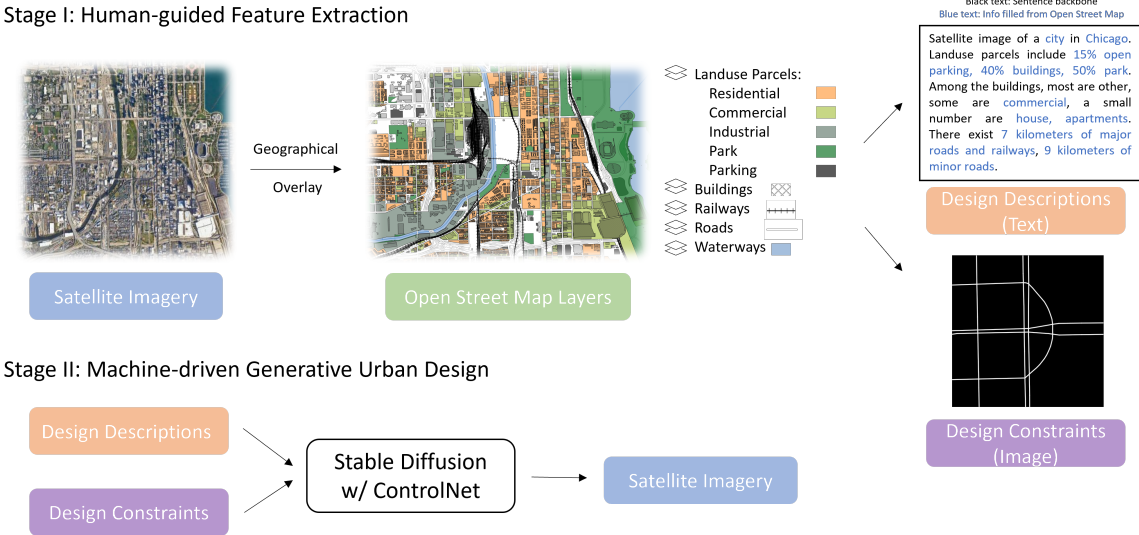


Figure 4-2: Two-stage model training

4.5.1 Feature Extraction

The first stage involves creating high-quality training data, which is crucial for effective model learning. For the model to generate new satellite imagery in response to descriptions and constraints, training pairs consisting of design constraints, design descriptions, and target satellite imagery need to be created. Figure 4-3 shows a few sample training pairs, followed by a detailed breakdown of each component.

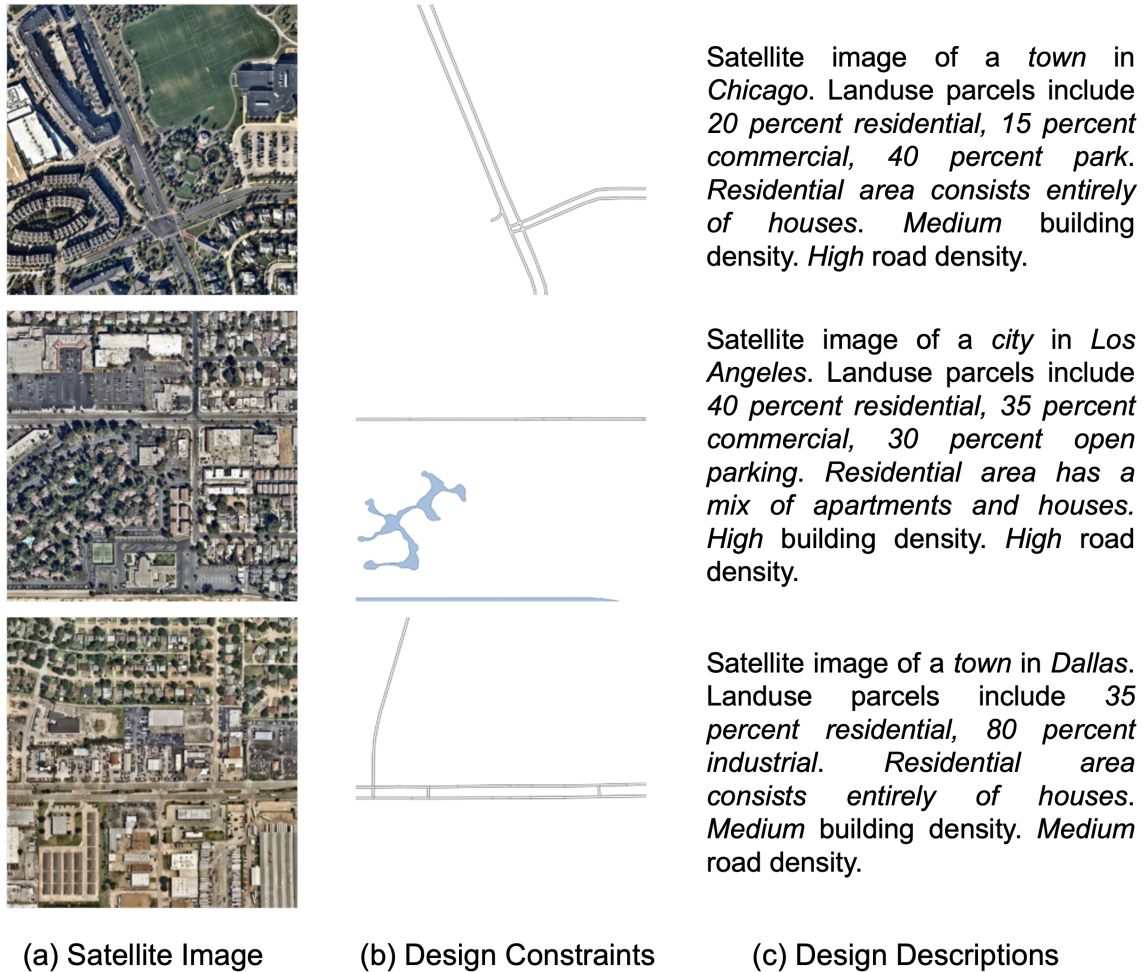


Figure 4-3: Sample Training Pairs

Target Satellite Imagery: We query the satellite imagery based on slippy map tile names [164] to obtain accurate coordinates for each tile. In this design scenario, the resolution was set to 450m x 450m, a design choice that echoes the 15 min city/neighborhood concept [163, 22, 106]. We aim to design a neighborhood that

is large enough to host a small mixed-use community with basic functionalities yet accessible within a 15-minute walk.

Design Constraints: Road infrastructure and natural environments were defined as constraints that will not be altered in the generation process. The ideal constraints would help frame the design space without being too restrictive. Therefore, the above-ground railways, major roads, and the waterways were selected as constraints. We extracted the railways, roads, and waterways layers from OpenStreetMap and used geoprocessing tools to extract the image for each tile so that the constraints would align with the true satellite imagery. Figure 4-4 shows the symbols for each constraint.



Figure 4-4: Symbols in the Design Constraint Image

Design Descriptions: The design descriptions have four components, each describing the setting, land use, residential type, and building footprint. We discuss how we came up with each component below. Texts in [] are information to be filled from geoprocessing.

1. **Setting:** Satellite image of a [city/town] in [metropolitan area name].

First, we want to orient ourselves. The city and the suburbs will look very different for each metropolitan area. The OpenStreetMap “places” layer indicates what type of “places” for each significant human settlement, such as city, suburb, town, village, etc. However, large pieces of urban areas are often missing in this layer, and the definitions for city/town/village are not rather fuzzy, leading to inconsistent labeling in different cities. Since we only consider the urban areas, we decided to only distinguish between city and town, with everything outside of the city municipality being labeled “town”.

2. **Land use:** land use parcels include [x percent of residential, x percent of com-

mercial, x percent of industrial, x percent of open parking, x percent of park, x percent of nature reserve].

Land use patterns are the most important design decisions in the process. Whether we are designing a residential area, a commercial area, or a mixed-use district will determine the appearance of the neighborhood. The OpenStreetMap “land use” layer describes the primary use of the land. We take six main categories and calculate the percent area each covers within a tile. Additionally, open parking spaces are a distinct land use feature on the satellite imagery that is not covered in the land use layer. This information can be found in the “traffic” layer and appended as one of the land use categories.

3. **Residential type:** Residential area consists [*entirely of houses/entirely of apartments/of a mix of apartments and houses*].

Different dwelling types will host very different population densities. Hence, the communities built around them will look very different. Therefore, we calculate the percentage of houses and apartments whenever buildings are labeled. Under the design scope of 450m x 450m, most of the areas only have one type. Instead of indicating the exact mix proportions, we only distinguish single-typed areas from mixed developments.

4. **Building area:** [*High/Medium/Low*] building coverage.

On the high level, the land use patterns loosely describe the main area functionalities. But, it is unknown exactly how much space is occupied by the buildings, as opposed to other roadside infrastructure or smaller facilities. The building density description complements the land use descriptions by stating exactly the percentage of land occupied by buildings. The “buildings” layer in OpenStreetMap delineates the outlines of buildings and houses. Ideally, we should further complement the land use descriptions with building types. However, the building type information is largely incomplete. Building density is measured by the percentage of land covered by buildings. Three categories were established based on natural breaks in the data found by the Fisher-Jenks algorithm: high

building density - [0.26,1]; medium building density - [0.10,0.26]; low building density - [0, 0.10).

4.5.2 Generative Urban Design

The second stage involves training a model to generate satellite imagery according to the description and constraints. Diffusion models were chosen due to their stable training properties, ease of scalability and fine-tuning, and their proven potential to extend to multi-modal inputs and outputs. We enable the multi-modal capability of diffusion models using ControlNet [186]. ControlNet is a framework created specifically to fine-tune diffusion models with custom conditions. It creates two copies of the blocks in the neural network: one “locked” copy and one “trainable” copy. The “locked” copy preserves the weights from a production-ready diffusion model, ensuring that high-quality images can be generated even from the beginning. The “trainable” copy gradually learns the custom condition. The final generation will be a weighted combination of both copies through a “zero-convolution” mechanism, which is a 1×1 convolutional layer with both weight and bias initialized to zeros.

Suppose a neural network block $\mathcal{F}(\cdot; \Theta)$, with trained parameters Θ , maps inputs x to outputs y : $y = \mathcal{F}(x; \Theta)$. Now we want to introduce custom condition c . Two instances of zero convolutions are applied on the custom condition $\mathcal{Z}(\cdot; \Theta_{z1})$, and the custom-conditioned output of the trainable copy $\mathcal{Z}(\cdot; \Theta_{z2})$. Then the ControlNet output y_c is

$$y_c = \mathcal{F}(x; \Theta) + \mathcal{Z}(\mathcal{F}(x + \mathcal{Z}(c; \Theta_{z1}); \Theta_c); \Theta_{z2}) \quad (4.1)$$

At the beginning, the weights of zero convolutions will be initialized to 0, and the output strictly comes from the production-ready diffusion model $\mathcal{F}(x; \Theta)$. As training progresses, the weights of the trainable copy increase, and the model will respond to the custom conditions. For the technical details of ControlNet, please refer to [186]. Stable diffusion, an open-sourced and production-ready diffusion model, is a robust and powerful backbone for fine-tuning. Therefore, we select stable diffusion as our

base model. The mathematical details about stable diffusion are in Appendix B.

4.6 Dataset

Data from three major U.S. cities was acquired: Los Angeles, Dallas, and Chicago. The geographical coverage of the dataset is shown in pink in Figure 4-5. To select useful design spaces, only the urban areas (shown in brown in Figure 4-5) defined by the U.S. Census [18] were selected. Furthermore, since OpenStreetMap is crowd-sourced data and can be incomplete in many ways, we filter to tiles with over 50% of the area being covered by major land-use patterns described above. After both filters, we had 18k, 7k, and 7k training samples for Chicago, Dallas, and Los Angeles, respectively. For all three cities, we first augment the training set by shifting the tiles in both the down and right directions by half of the tile width. In addition, we further augment the Dallas and Los Angeles datasets by shifting the tiles by half of the tile width right and down separately, thus quadrupling the sample size. As a result, we had 36k, 28k, and 28k samples for Chicago, Dallas, and Los Angeles, respectively.

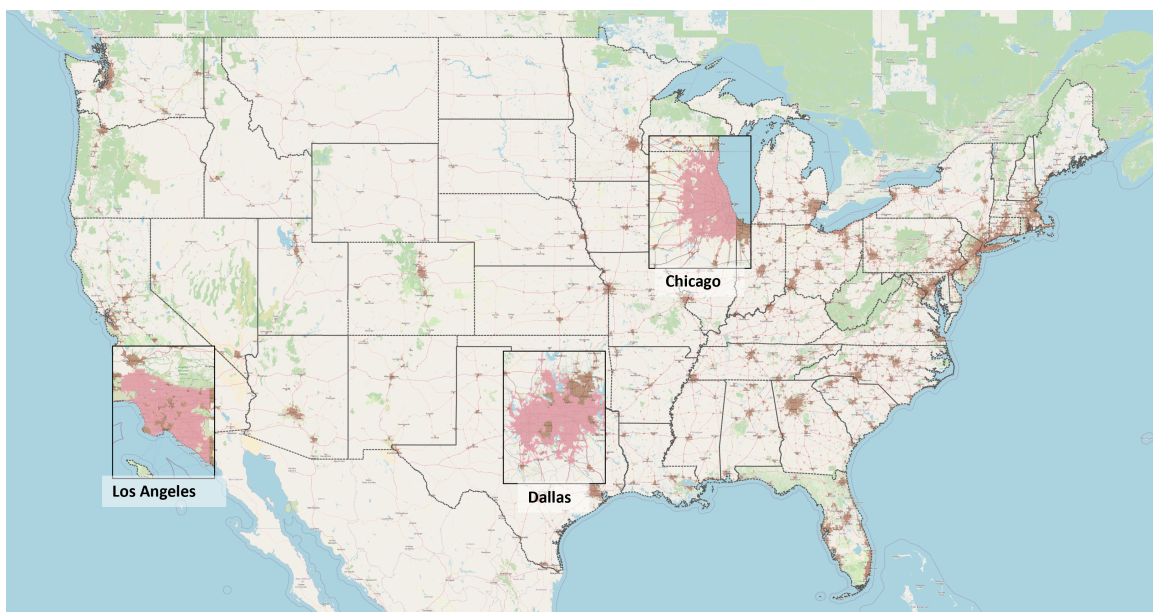


Figure 4-5: Dataset Coverage

4.7 Model Evaluation

A user study was conducted to receive evaluation and feedback for the trained model. The user study has two parts: scoring the selection. Appendix C includes a sample question from both parts. In the scoring part, users are asked to rate an image with a score between 1-5 on the images’ consistency with the described land use, on the degree to which existing infrastructure and natural environment are respected, and on the realism of the images. Each user is randomly presented with either the real or the generated image. In the selection part, a design description and a constraint image are presented alongside two satellite images, one real and one generated. The user is asked, “Which image reveals an urban environment closer to the language description?” While the first part provides the breakdown of each of the goals for the model, the second part of the study reveals user preferences in real scenarios. 20 mixed-use (having 3 or more land use types) neighborhoods are selected for this evaluation. Part 1 and Part 2 contain the same pool of images.

The generated images have successfully learned the features in real images. Table 4.1 tabulates the scores. The generated images are slightly worse regarding matching land use patterns and better regarding conforming to the constraints. The generated images received lower scores on realism due to the typical higher saturation of the generated images, and the trees being oil painting style. In addition, from Part 2 of the user study, out of the 20 pairs, 12 generated aerial images are selected over the real images as better revealing the description.

	Land use	Constraint	Realism
Real	3.73	3.68	3.87
Generated	3.39	3.94	3.06
Difference	-0.34	+0.26	-0.81

Table 4.1: User study scores of generated and real satellite images (Min score:1, Max score:5)

4.8 Model Demonstration

In this section, we qualitatively demonstrate the model’s capabilities. First, we illustrate that the goals outlined in Section 4.4 are achieved. Second, an ablation study is performed on the components of the design description. Lastly, we present some thought experiments that could be performed using this model.

4.8.1 Goals

Having Control over AI-generated Land Use Patterns: First, we create images based on real descriptions taken from the dataset, and we compare these generated images with the actual images. While they don’t have to be identical, we should notice resemblances and be able to account for the differences. The figures in Table 4.2 show one sample from each city. The first and second columns show the design description and constraint, and the third column shows the true satellite image, with the land use annotation taken from OpenStreetMap. The final column displays the generated image. Upon qualitative examination, we can identify the various land use parcels within this image, which we have marked directly on the image. Although the spatial placements of the land use parcels are different for the true and the generated images, the types and the approximate proportions are similar in all three cases.



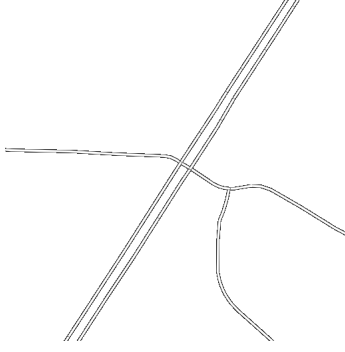


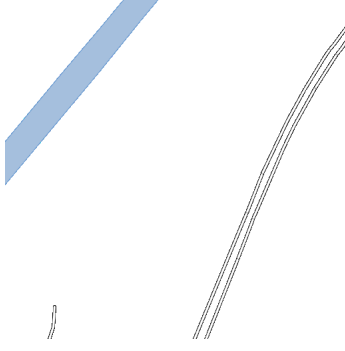


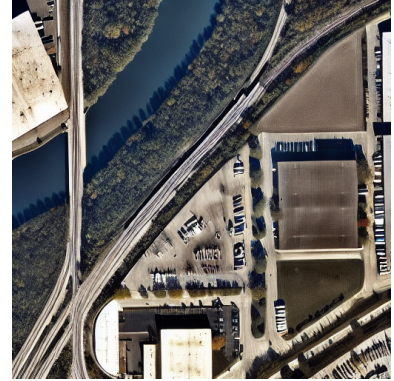
Design Description	Design Constraint	True Satellite Image	Generated Satellite Image
<p>Satellite image of a city in Chicago. Land use parcels include 40 percent residential, 15 percent industrial, 15 percent commercial, 10 percent park, 5 percent open parking. Residential area consists entirely of houses. Medium building coverage.</p>			
<p>Satellite image of a town in Dallas. Land use parcels include 30 percent residential, 25 percent industrial, 20 percent commercial, 20 percent farmland. Residential area consists entirely of houses. Medium building coverage.</p>			
<p>Satellite image of a town in Los Angeles. Land use parcels include 65 percent residential, 20 percent park. Residential area has a mix of apartments and houses. Medium building coverage.</p>			
<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="border: 1px solid orange; padding: 2px 5px;">Residential</div> <div style="border: 1px solid blue; padding: 2px 5px;">Commercial</div> <div style="border: 1px solid green; padding: 2px 5px;">Park</div> <div style="border: 1px solid yellow; padding: 2px 5px;">Industrial</div> <div style="border: 1px dashed blue; padding: 2px 5px;">Open parking</div> <div style="border: 1px solid darkgreen; padding: 2px 5px;">Farmland</div> </div>			

Table 4.2: True vs. generated satellite imagery

Additionally, we convert existing land use patterns into new ones to demonstrate the model’s understanding of various land use patterns. The figures in Table 4.3 show different land use alternatives for the three places in Table 4.2. Among the alternatives, we show the model’s understanding of different proportions of residential, industrial, commercial, park, nature reserve, and open parking spaces. *Residential* areas are characterized by large, dense connected buildings with green space and minor roads. The roads have are a mix of curved and straight grids. *Commercial* areas are characterized by larger buildings, with larger spacing. Often accompanied with small amounts of other amenities such as parking lots and green space. *Industrial* areas are often large and isolated. The buildings have large footprint, and usually there are no other amenities of close proximity. *Open parking* areas are small, open spaces that are often attached to a commercial or industrial building. From an aerial perspective, some cars are usually visible. *Parks* are smaller green space that often on river banks, and/or surrounded by residential or commercial areas. There are often patterns of small amenity buildings, or shapes of sports fields. *Nature reserve* are isolated, large areas of green space and waterways with a lot of tree covers. *Farmland*, similarly, are large pieces of land, with no visible artifacts from a satellite image.

Producing alternate designs for the same inputs: In addition to responding to prompts, Figures in Table 4.4 demonstrate that the model can generate multiple designs given the same prompt. Each row of Table 4.4 displays three alternative designs in response to the description and constraints for each instance in Table 4.2. Qualitatively, the spatial layout and shapes of the land use parcels are different in each set of designs, while the components and the proportions are similar.

Base Description: Satellite image of a city in Chicago. Land use parcels include $\{ \}$. Medium building coverage.



Base Description: Satellite image of a town in Dallas. Land use parcels include $\{ \}$. Medium building coverage.



Base Description: Satellite image of a town in Los Angeles. Land use parcels include $\{ \}$. Medium building coverage.



60% residential, 20% park

60% commercial, 20% open parking

30% industrial, 25% nature reserve, 25% farmland

Table 4.3: Generations over different land use combinations

Design Description	Design 1	Design 2	Design 3
<p>Satellite image of a city in Chicago. Land use parcels include 40 percent residential, 15 percent industrial, 15 percent commercial, 10 percent park, 5 percent open parking. Residential area consists entirely of houses. Medium building coverage.</p>			
<p>Satellite image of a town in Dallas. Land use parcels include 30 percent residential, 25 percent industrial, 20 percent commercial, 20 percent farmland. Residential area consists entirely of houses. Medium building coverage.</p>			
<p>Satellite image of a town in Los Angeles. Land use parcels include 65 percent residential, 20 percent park. Residential area has a mix of apartments and houses. Medium building coverage.</p>			

Table 4.4: Balance between creativity and design descriptions

Respecting Existing Infrastructure and Natural Environment: Regardless of the type, shape, and area of the constraints, the model will respect these given constraints. Constraints in this study include major railways, roadways, and waterways. The figures in Table 4.5 show the model’s generation under different combinations of constraints.

Additionally, we show the transformation in generations when different types of constraints are added in Table 4.6. The common design description used is “Satellite image of a city in Chicago. Landuse parcels include 35 percent residential, 20 percent commercial, 5 percent open parking. Residential area has a mix of apartments and houses. Dense building coverage. ” Starting from a blank image, we add the waterways, railways, and roads sequentially as we move down the rows of Table 4.6. The generations, correspondingly, started from a typical grid structure, and sequentially incorporated the constraints into the generations.

Learn and Apply Local Urban Textures in Different Cities: Given the same constraint and description, the figures in Table 4.7 show that the generations in three cities look distinctly different. Each row in Table 4.7 used the same description and constraint, except for the city name (made *italic* in the description). One distinct feature to notice is that green space and trees are often associated with houses in Chicago and Dallas, but not in Los Angeles. Additionally, the footprints of individual houses in Chicago are generally smaller compared to Dallas and Los Angeles. Houses in Los Angeles often have swimming pools, marked by light blue spots.

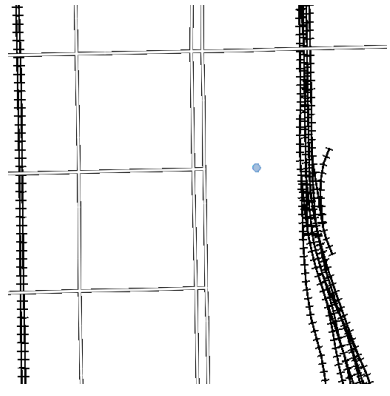
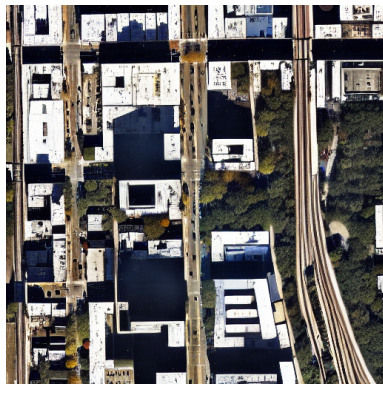
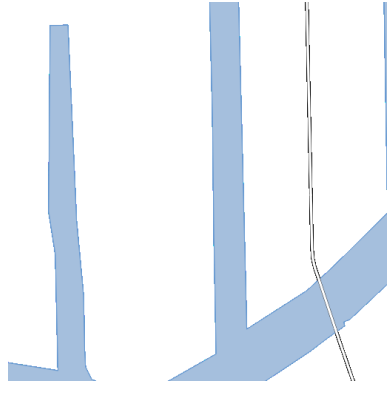

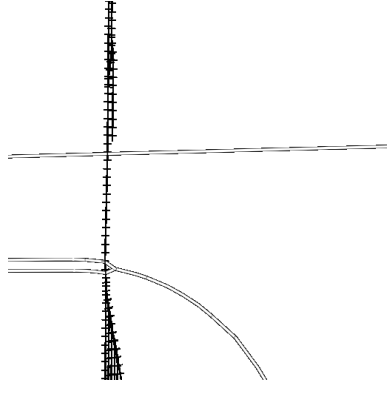

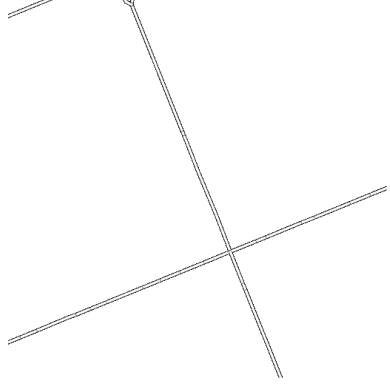
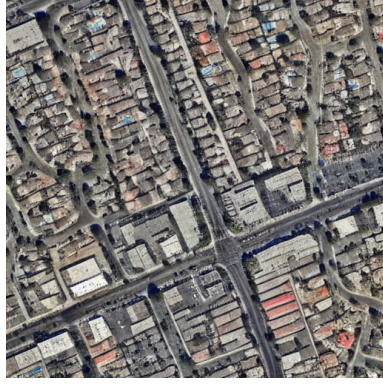
Constraints	Constraint Image	Design
Railways, roads		
Waterways, roads		
Railways, roads		
Roads		

Table 4.5: Adherence to different types of design constraints


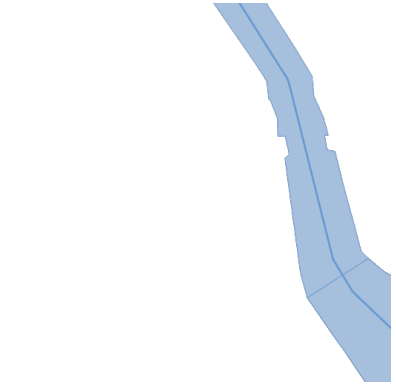

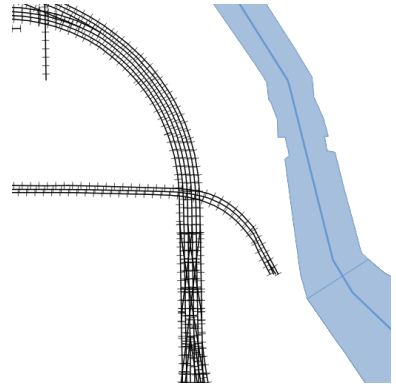

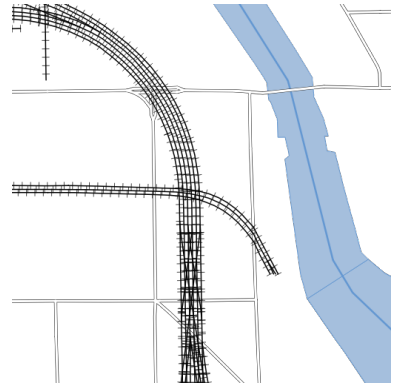

Constraints	Constraint Image	Design
No Constraint		
Add Waterways		
Add Railways		
Add Roads		

Table 4.6: Sequentially adding design constraints to generations

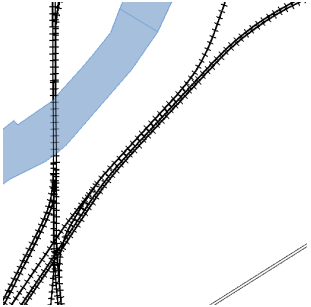


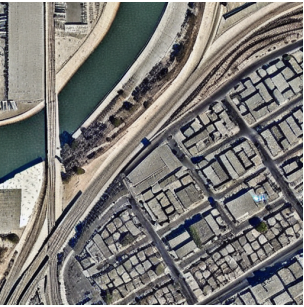
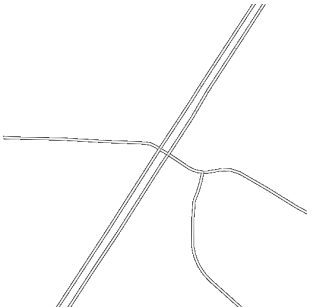



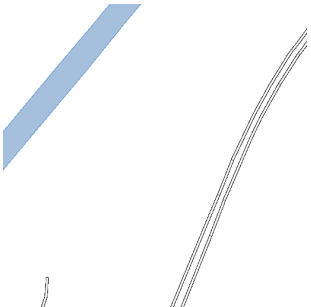



Design Description	Design Constraint	Chicago	Dallas	Los Angeles
Satellite image of a city in [city name]. Land use parcels include 40 percent residential, 15 percent industrial, 15 percent commercial, 10 percent park, 5 percent open parking. Residential area consists entirely of houses. Medium building coverage.				
Satellite image of a town in [city name]. Land use parcels include 30 percent residential, 25 percent industrial, 20 percent commercial, 20 percent farmland. Residential area consists entirely of houses. Medium building coverage.				
Satellite image of a town in [city name]. Land use parcels include 65 percent residential, 20 percent park. Residential area has a mix of apartments and houses. Medium building coverage.				

Table 4.7: Local urban textures in different cities

4.8.2 Ablation Study

In this section, the effect of each component of the description is isolated. Since land use patterns and city names were examined in detail in Section ??, they will be fixed in this part. The figures in Table 4.8 use the same base description, with only one component changed for each image. For example, the description used to generate the image on row #2 and column “All houses” is: “Satellite image of a city in Chicago. Land use parcels include 40 percent residential, 20 percent commercial, 20 percent park. Residential area consists of all houses. Sparse building coverage.”.

City/Town: A distinct difference can be observed between the urban and suburban environments. In general, the buildings are smaller in the city and the trees are scattered instead of a whole area covered entirely by trees.

Houses/Apartments: A transition in building footprints can be observed. The left side of the image is created as residential. In the “all houses” scenario, houses form a uniform grid. With the introduction of apartments, larger and more irregularly shaped buildings started to appear.

Building coverage: Going from left to right, the area covered by buildings is increasing. The model learns to create larger and denser buildings. For example, in the dense scenario, we do not see trees on the left part of the image, and the houses are very densely packed.

4.9 Discussion, Limitations, and Future Work

Our model demonstrates a controlled synthesis of satellite images for urban design, both in quantitative and qualitative terms. However, this study is an early attempt at developing a suite of GenAI-based models for urban design. In an effort to maximize the capabilities of generative AI tools and guide subsequent research, we conducted interviews with several scholars and practitioners. Their insights on using generative AI tools and suggestions for essential features to be incorporated in future iterations were invaluable. In this section, we share a few lessons learned during the development of this chapter, recognize some model limitations, and suggest potential directions for









Base Description			
Satellite image of a city in Chicago. Land use parcels include 40 percent residential, 20 percent commercial, 20 percent park. Sparse building coverage.			
#1			
	City	Town	
#2			
	All houses	All apartments	A mix of houses and apartments
#3			
	Sparse building coverage	Medium building coverage	Dense building coverage

Table 4.8: Ablation study of design descriptions

future research.

Potential use cases: The model cannot be directly applied to planning projects

in its present form. However, despite being an initial iteration of a more comprehensive version, the model does possess certain value-adds to the industry.

First, the model allows for rapid visualizations of conceptual designs. Planners and the public can pose “what-if” questions regarding their neighborhoods and receive visualizations in almost real-time. Valid queries include “What if we remove this major highway ramp out of our neighborhood?”, “What if we turn the residential area into commercial?”, “What if we build a park in the area?”, etc. While the AI tool doesn’t offer precise designs, it does furnish a bird’s-eye perspective to aid in intuitive understanding.

Second, it can uncover associations and styles that are harder to articulate. The model learns styles from different cities. By contrasting visualizations across these cities, one can draw inspiration for elements not initially considered in the design constraints and prompts.

Third, it helps non-professionals to engage with and explore professional concepts around urban design. This tool enables those without professional experience to delve into urban design effortlessly. It has the potential to inspire public involvement and creativity in local planning initiatives.

Data limitation: In this section, we discuss one limitation incurred by the data. Modeling limitations and future work will be discussed in the next section. The proportions of land use patterns being generated in the images are approximations. To resolve the common challenge of insufficient data, we proposed using a widely available and standardized data source - OpenStreetMaps- to curate labeled training data. The advantage is clear - we solve the problem of insufficient amount of training data, and our model can be easily generalized due to standardization. The disadvantage is that OpenStreetMaps is crowd-sourced. The data quality varies between regions, and much pre-processing is needed. The land use labels are missing in some places, resulting in imprecise descriptions of the land use proportions. Therefore, the learned representations are approximate.

Future research: Despite the current advancements in GenAI, it does not fully consider real-world applications’ complexities. Several recurring themes emerged in

our interviews. Tackling these issues presents significant challenges, whether from a technological or data collection standpoint. As such, we underscore them as key areas for future research.

First, satellite image as the sole output is not informative enough. While satellite imagery provides a comprehensive overview of an area, its utility is limited, particularly due to the height information. This limitation becomes apparent in complex urban layouts, where buildings may house commercial spaces on the ground floor and residential units above. From the bird's eye view, a warehouse, which is industrial, and a large grocery supercenter, which is commercial/retail, might look the same. Additionally, relying solely on satellite images precludes thorough assessments of key factors like economic impact, feasibility, sustainability, and walkability, all of which are vital in real urban design scenarios. Accurate knowledge of the generated building boundaries, intricate road networks, and any bicycle or public transit infrastructure is essential to conducting these assessments. Therefore, future research should produce a comprehensive set of outputs beyond just satellite imagery, such as 3D representations, street views, and land use layouts.

Second, it is important to improve the details of the inputs as much as we focus on improving the suite of the outputs. The current descriptors are direct descriptions of approximate land use patterns. In actual planning practices, more nuanced descriptors are used. For instance, zoning data offers a nuanced perspective on land use, including specifics like building setbacks and floor-to-area ratios. However, the challenge lies in the fact that zoning regulations vary significantly across cities, posing considerable difficulties in data collection and processing. Additionally, the topography of an area has profound implications on plan feasibility, influencing everything from building designs to infrastructure development. Therefore, incorporating terrain data into the model would be a significant step toward creating more realistic and practical urban design tools.

Third, the descriptors are direct descriptions of land use patterns with no sociodemographic associations considered. Attempts were made to use social descriptors such as high-income, high-walkability, transit-oriented, and well-educated in the cur-

rent architecture, but it cannot learn from them. This limitation may stem from the challenge of establishing clear links between these descriptors and land use patterns, or it could be due to our dataset not being sufficiently comprehensive for the model to recognize these more nuanced associations. Future research should consider incorporating social descriptors into the descriptions to inspire designs with desired characteristics.

Fourth, the framework could benefit from image editing capabilities. Currently, a brand new generation is done each time the description or the constraint is modified. We cannot choose which parts of the design we want to keep or change. As a first step, the introduction of 'masks' could be beneficial. These masks would allow for the exclusion of certain areas from changes, maintaining the integrity of specific parts of a town. To take it further, we should be able to edit the style of certain parts of the image while keeping the content intact.

Fifth, value judgments are not introduced in the current framework. While human planners possess intuitive and experiential understandings of what constitutes good design, these criteria are not explicitly conveyed to the AI system. Our current approach focuses on learning the existing urban patterns – the status quo – without embedding specific value judgments into the process. Consequently, the AI operates without distinguishing between the desirable and undesirable aspects of current urban layouts. This absence of value judgments means that the AI might replicate existing designs indiscriminately without consideration of their qualitative implications. Urban development often aims to mirror the present and improve upon it, addressing its strengths and shortcomings. To align the AI's output with these developmental goals, it's necessary to integrate evaluative criteria into the model. This would enable the AI to differentiate between the positive and negative elements of existing urban designs and generate plans that reflect current realities and aspire to enhance them.

Sixth, within a single city, a diverse range of styles may coexist due to historical reasons or different planning considerations. However, our current model assumes that each city has a uniform style. This approach overlooks the reality that each neighborhood and building can embody a distinct style. Acknowledging and incor-

porating this diversity would greatly enhance the model’s accuracy and relevance in urban design contexts, allowing it to represent the unique architectural character of each part of the city more faithfully.

4.10 Conclusion

The iterative nature of urban design favors tools that streamline the design process, enhance communication, and provide quick feedback. Artificial Intelligence (AI) has recently shown promising urban design capabilities, such as generating data-driven insights, assessing and improving plan performance, and crafting visualizations. Despite AI’s significant potential to revolutionize urban design, practical implementation faces numerous challenges. Contemporary research is primarily geared towards narrowly defined, tightly controlled tasks. Urban design, with its variety and complexity, does not lend itself easily to simple parameterization. While it’s widely agreed that the future of urban design will involve a synergy between humans and machines, a detailed framework outlining this collaboration is lacking. Additionally, generative AI’s reliance on substantial data volumes poses a problem, as acquiring labeled data in urban contexts is costly and thus limited. This study addresses these issues by proposing and implementing a novel human-machine collaboration framework in urban design. This framework is brought to life through a model trained on data automatically labeled from available resources, offering a novel approach to integrating AI in urban design and bridging the current gap between theoretical potential and practical application.

The developed model can produce satellite imagery from design constraints in the form of images and descriptions in the form of text. It enables human-guided control over AI-generated land use patterns, offering creativity to generate varied designs under the same constraint and description. The model also consistently respects various types of constraints and adeptly incorporates local textures from different cities in the designs. The trained model could be used for rapid visualizations of conceptual designs, uncovering implicit associations and styles, and assisting non-

professionals to engage with and explore professional urban design concepts.

We are in the early stages of generative AI development. While the model has shown promising results in qualitative and quantitative assessments, many limitations remain, leading to six key areas for future research. First, a more comprehensive range of outputs must be developed to enhance urban plan representation and evaluation. Second, incorporating more details in inputs can contribute to the realism and practicality of the plans. Third, learning social descriptors could inspire designs with desired social characteristics. Fourth, allowing editing can promote the usability of the tool. Fifth, incorporating value judgments into the generative process is essential for designing beyond the current urban landscapes. Finally, a greater variety of styles within a single city should be considered.

Chapter 5

Conclusion

This dissertation examines three emerging issues in urban computing: uncertainty quantification, data fusion, and generative urban design. The motivation and conclusions of each study are discussed in detail in each chapter. To conclude this dissertation, the high-level contributions of each study and the connections between the chapters are summarized, followed by a discussion of many opportunities for future research.

5.1 Summary

First, this dissertation identifies the lack of uncertainty quantification in short-term spatiotemporal demand prediction literature and develops a probabilistic graph neural network framework to focus not only on the mean but also on the distribution of the output. Prob-GNNs can accurately predict ridership uncertainty, even under significant domain shifts such as COVID-19. Prob-GNNs play a significant role from both the theoretical and practical perspectives. Theoretically, Prob-GNNs provide a probabilistic view of travel demand and describe distributional characteristics. Practically, demand predictions with uncertainty estimates allow more efficient resource allocation in downstream tasks. For example, resources can be allocated to places with more certain demand. Additionally, spatiotemporal uncertainty patterns inform operators of potential anomalies, disruptions, and bottlenecks in the system.

Second, this dissertation develops deep hybrid models (DHM) to combine numeric and imagery data for travel behavior analysis. DHM aims to enrich the family of hybrid demand models using deep architectures. DHM demonstrated that urban imagery and sociodemographics contain complementary information because DHM outperforms benchmarks using only one of the two data sources in both the aggregate and disaggregate mode choice prediction tasks. In addition, DHM constructs a latent space capable of generating new satellite images that do not exist in reality and computing the corresponding travel behavior and economic information, such as substitution patterns and social welfare. DHM is a versatile framework that combines and creates associations between different types of numeric, unstructured data, and output variables. But DHM is more than a data fusion framework. DHM is a data fusion framework with a generative view. Through the latent space, DHM enables the generation of unstructured data and the association of such generated data with numeric data and output variables. As a result, DHM is a major move towards a comprehensive embedding of relevant urban information which offers opportunities for future research on the predictive capabilities and latent space interpretation to deepen the understanding of connections between different urban elements.

Lastly, this dissertation creates a human-machine collaboration framework for generative urban design and showcases a model trained to generate satellite imagery from a land use text description and a constraint image depicting the built and natural environment not to be altered. The trained model can generate high-fidelity, realistic satellite images while retaining control over the land use patterns in generated images with natural language descriptions, producing alternate designs with the same inputs, respecting the built and natural environment, and learning and applying local contexts from different cities. The trained model could be used for rapid visualizations of conceptual designs, uncovering implicit associations and styles, and assisting non-professionals to engage with and explore professional urban design concepts.

This dissertation signals the rapidly advancing capabilities of deep learning in processing more complex data sources. In the uncertainty quantification chapter, all inputs and outputs are numerical. We aim to draw attention to uncertainty quantifi-

cation in spatiotemporal demand prediction using deep learning, which is well-studied in traditional time series and spatial lag and error models. The data fusion chapter introduces unstructured data to enrich the traditional choice models. The resulting deep hybrid model framework enhances the predictive power of traditional choice models and enables additional capabilities to conduct associative analysis between urban imagery, sociodemographics, and travel behavior. The generative urban design chapter completely operates in the space of unstructured data. Recent advances in computational power and large models facilitate the interface between humans and machines through human-friendly images and natural language. This advancement greatly reduced the complexities of applying deep learning techniques, making them more accessible for addressing various aspects of urban development.

The three chapters identify three themes for approaching urban challenges using deep learning: mimicking, enriching, and surpassing traditional models. The uncertainty quantification chapter calls the attention of deep learning models to the overlooked uncertainty quantification, which is well-studied in traditional models. The data fusion chapter aims to enrich the traditional choice models to offer better predictive power and to allow for a deeper understanding of travel behavior and urban imagery. The generative urban design chapter empowers new urban computing tasks: directly generating satellite imagery for urban designs was unheard of before the advent of deep learning. This dissertation provides ways of thinking about harnessing the power of deep learning. As researchers become more adept, we can anticipate further advancements in urban computing with more creative uses of deep learning.

In summary, the intersection of deep learning and urban computing marks a significant stride toward solving urban challenges with more advanced, efficient, and innovative methods. As this field continues to evolve, it will undoubtedly play a critical role in shaping our cities' future and quality of life.

5.2 Future Research

This dissertation takes one step in each area of interest, while many unexplored research topics remain. Some follow-up research has already been conducted within the JTL mobility lab from which this dissertation stems. This section identifies such studies and presents other examples of potential topics along each line of research.

5.2.1 Uncertainty Quantification

Chapter 2 identified that uncertainty quantification has both theoretical and practical significance. Theoretically, we want to develop a more correct and comprehensive understanding of travel demand; practically, predictive uncertainty can help optimize downstream decision-making. Therefore, we identify future works from both theoretical and practical perspectives.

In Chapter 2, we tested six probabilistic assumptions and two deterministic assumptions. As a direct extension, testing our framework under varying data scales and contexts is important to corroborate the above conclusions and better inform policy-making. For example, one scenario that requires special attention is extreme sparsity, which is investigated concurrently. Zero-inflated negative binomial and Tweedie distribution best predicted the overly sparse and dispersed ridesharing demand [194, 69]. Other works have used this framework for incident prediction [40] and data imputation [161].

Prob-GNN is a parametric framework for uncertainty quantification. In addition to the parametric framework, the uncertainty quantification literature has three other major frameworks: Bayesian, non-parametric, and calibration methods. Therefore, a more detailed investigation into Bayesian, non-parametric, and other calibration methods is needed regarding the applicability to various fields of urban computing. One follow-up study has shown that applying calibration to sparse and asymmetrically distributed prediction problems reduces calibration error by 20% [192]. Therefore, it is important to compare the different uncertainty quantification frameworks in different uncertainty scenarios.

The most important practical implication for uncertainty quantification is to more accurately inform downstream decision-making so that resources can be allocated more efficiently. In the current optimization literature, the uncertainty set is often directly derived from historical observations or defined arbitrarily. A follow-up study compared the effectiveness of vehicle rebalancing based on different demand prediction algorithms, including true demand, historical average, point prediction, and data-driven uncertainty interval [49]. The study has shown that if demand can be predicted accurately, rebalancing according to the predicted demand from the point prediction method is good. However, when the demand is volatile and hard to predict, rebalancing according to the uncertainty intervals yields better results. Future research should further explore the interaction between the upstream prediction task and the downstream decision-making (e.g. rebalancing, dispatching, and dynamic tolling). For example, exploring the optimal prediction interval width, the effect of predictive bias, and the problems associated with predictions.

Furthermore, the current literature still treats the prediction and optimization problems as separate processes. It would be optimal if an integrated framework for both tasks could be created. Ultimately, the optimization outcomes directly influence rider experience and service reliability. Existing efforts can be categorized into two categories: formulating simple machine learning prediction methods into optimization [15] and building end-to-end machine learning pipelines by learning the optimization outcomes [81, 12, 116]. Despite the efforts, the integration often faces challenges due to the complexity of both the machine learning and the optimization problems. This research area could benefit immensely from interdisciplinary methodological innovations in machine learning and optimization.

5.2.2 Data Fusion: Deep Hybrid Models

The deep hybrid model framework aims to leverage the unique capacity of deep learning to process unstructured data. The DHM serves two purposes: representation learning for prediction and image regeneration for image-based story-telling. Future work should consider advancing either or both fronts to further develop and enrich

the framework.

The framework consists of four major components: unstructured data (input), the outcome variable, the mixing operator, and the behavioral predictor. Each component of the DHM framework can be substituted for another context and presents major opportunities to deepen the understanding of the unstructured data and outcome variables. Aside from satellite imagery, the road network also plays a role in determining the mode of transportation people choose. A working paper that models travel behavior using the road network is being prepared. Additionally, with the emergence of large language models, there are also works incorporating natural language descriptions to enrich further unstructured datasets used in travel behavior modeling. In addition to travel behavior, sociodemographics are often considered exogenous variables for energy consumption, health, and air pollution, so future research could analyze these outcome variables using the DHM framework. Finally, alternative mixing operators could be explored to construct latent spaces for better representation learning or image generation. For instance, a working paper is being prepared to explore discrete latent spaces with VQGAN.

The framework is inevitably associated with the limitations and challenges of adopting complex neural networks. Overcoming these limitations and challenges is crucial to advancing both deep learning and its application in urban computing. Here, we highlight a few themes: non-identification, transfer learning, and model interpretation. First, deep learning can effectively minimize errors, but they always converge to local minima. This non-identification issue is still a pending challenge in deep learning research. Although it is not necessarily a major problem in learning tasks, it is ambiguous how to guarantee stability and robustness in the latent space. Second, The DHMs achieve relatively high predictive performance, but their transferability could be limited. It is unclear whether the models can perform well when trained in one context and used in another. If proven to be transferable, these models could greatly help travel behavior modeling in places where surveys are expensive and difficult to conduct. Thus, developing methods to evaluate and enhance model transferability would be highly valuable. Third, the economic information computed from the gen-

erated satellite imagery is a numerical approximation, unlike the analytical solutions from the classical parameter-based discrete choice models. The latent space in DHMs is no longer constrained to a few latent variables, limiting its interpretability in the classical parameter-based sense. To improve the reliability and trustworthiness of our models, we should explore alternative methods for interpreting the latent space. This could potentially provide valuable insights into the workings of deep learning models and help us better understand the association between sociodemographics, urban imagery, and travel behavior.

5.2.3 Generative Urban Design

Since we are in the early stages of generative urban design development, Section 4.9 is dedicated to discussing the potential use cases, limitations, and future research directions. This section summarizes the considerations in Table 5.1 for easier reference.

Limitation	Future work
Satellite image as the sole output has limited expressivity.	Developing a more comprehensive range of outputs such as land use layouts and street view.
The inputs are restricted to land use descriptions only	Including detailed design specifications such as zoning, setback, and floor-to-area ratios.
	Learning social descriptors such as elderly-friendly, high-income, and well-educated.
Designs need to be regenerated every time the description or the constraint changes.	Building image editing capabilities (e.g. masking, style editing.)
Designs cannot be quantitatively evaluated.	Incorporating value judgments into the generative process.
A city is assumed to have a single style.	Acknowledging a variety of styles within each city.

Table 5.1: Limitations and Future Work of Generative Urban Design

Appendix A

Additional Details of Deep Hybrid Models

A.1 Specific model design in disaggregate analysis

To simplify the notation used in the manuscript, we only distinguish between imagery latent variables and numeric variables in our discussion. In fact, there are three types of inputs to the behavioral predictors: imagery latent variables, sociodemographics, and alternative-specific variables. Both sociodemographics and alternative-specific variables are numeric but enter the utility function slightly differently. In this paper, we used z_n to represent the imagery latent space, x^{sd} and x^{alt} represent the sociodemographic and alternative-specific attributes, respectively. Combining all three types of attributes, we formulate the utility function as follows:

$$V_{nk} = \beta_k^{im'} z_n + \beta_k^{sd'} x_n^{sd} + \beta^{alt'} x_{nk}^{alt} \quad (\text{A.1})$$

where $\beta_k^{im'}$, $\beta_k^{sd'}$, and $\beta^{alt'}$ are vector coefficients of urban imagery latent space, sociodemographics, and alternative-specific attributes, respectively. We assume that the alternative-specific variables share the same coefficients in the utility of all alternatives. Both imagery and sociodemographics are not alternative-specific. Therefore, the coefficients $\beta_k^{im'}$, $\beta_k^{sd'}$ are alternative-specific to distinguish the effect these vari-

ables have on different alternatives.

Section 3.3.4 also discussed the directional probability derivatives for the imagery latent variable z_n . We could also calculate probability derivatives for the other two types of variables. Both sociodemographic and alternative-specific variables are standard numeric variables, and many references exist on how to get the probability derivatives. We only summarize the results below.

For alternative-specific attributes, the probability derivative, which is the change in the probability that a trip n uses alternative k with respect to changes in variable i is simply

$$\frac{\partial P_{nk}}{\partial x_{nki}^{alt}} = \frac{\partial V_{nk}}{\partial x_{nki}^{alt}} P_{nk}(1 - P_{nk}) = \beta_i^{alt} P_{nk}(1 - P_{nk}) \quad (\text{A.2})$$

Sociodemographic variables enter the utility of all alternatives with different β parameters. Therefore, the change in probability that a trip n uses alternative k with respect to changes in variable i is

$$\frac{\partial P_{nk}}{\partial x_{ni}^{sd}} = \sum_{k'} \frac{\partial P_{nk}}{\partial V_{nk'}} \frac{\partial V_{nk'}}{\partial x_{ni}^{sd}} = \beta_{ki}^{sd} P_{nk}(1 - P_{nk}) - \sum_{k'} \beta_{k'i}^{sd} P_{nk} P_{nk'} \quad (\text{A.3})$$

We can compute a directional gradient of choice probabilities regarding the imagery latent space, which resembles but differs from the marginal effects β in the classical demand modeling. The formula of the directional gradient is

$$\nabla_u P_{nk}(z) = u \cdot \nabla P_{nk}(z) = \sum_{k'} \frac{\partial P_{nk}}{\partial V_{nk'}} \nabla_u V_{nk'}(z_n) = u \cdot (\beta_k^{im} P_{nk}(1 - P_{nk}) - \sum_{k'} \beta_{k'}^{im} P_{nk} P_{nk'}) \quad (\text{A.4})$$

A.2 Effects of mixing and sparsity hyperparameters on prediction and image reconstruction

The mixing hyperparameter λ affects the performances in behavioral predictors by controlling the degree of information mix in the latent space. The degree of information mix can be understood by the trade-off between retaining the richness in

image reconstruction and encoding sociodemographic information. Table A.1 shows the predictive performance under different λ values. Since $\lambda = 0$ represents an autoencoder without sociodemographic information, the non-zero λ values in Table A.1 imply different degrees of information mixing. The improvement of performance can be quite significant by optimizing λ . The mixing hyperparameter should not be set too small or too large. Among the five values tested, $\lambda = 0.7$ and 0.9 achieve the best performance in Panel 1, while $\lambda = 0.5$ and 0.7 achieve the best performance in Panels 2 and 3. Overall, $\lambda = 0.7$ achieves near-optimal performance across all panels, and therefore, it is used in the subsequent analyses. This observation corroborates the claim that the two data sources are complementary and achieve the highest predictive performance when properly mixed.

Table A.1: Effects of the mixing hyperparameter λ in Model 4

The mixing hyperparameter λ	0.1	0.3	0.5	0.7	0.9
<i>Panel 1: Aggregate Mode Choice - Linear Regression</i>					
Auto (R^2)	0.621/0.593	0.635/0.592	0.688/0.633	0.676/0.640	0.688/0.637
Active (R^2)	0.501/0.474	0.549/0.480	0.582/0.528	0.565/0.524	0.589/0.533
PT (R^2)	0.463/0.444	0.514/0.473	0.538/0.473	0.520/0.483	0.519/0.488
<i>Panel 2: Aggregate Mode Choice - Multinomial Regression</i>					
KL Loss	0.139/0.147	0.135/0.148	0.110/0.141	0.101/0.140	0.104/0.145
Auto (R^2)	0.606/0.581	0.607/0.559	0.706/0.605	0.737/0.604	0.734/0.590
Active (R^2)	0.484/0.433	0.493/0.435	0.630/0.472	0.657/0.485	0.668/0.469
PT (R^2)	0.437/0.393	0.481/0.406	0.608/0.421	0.674/0.421	0.650/0.363
<i>Panel 3: Disaggregate Mode Choice</i>					
CE Loss	0.376/0.402	0.378/0.412	0.371/0.406	0.373/0.404	0.461/0.467
Accuracy	0.870/0.858	0.870/0.856	0.872/0.861	0.871/0.856	0.843/0.831

The reconstruction quality between different λ 's does not differ much because of the various image quality enhancement terms (GAN, KL, LPIPS) in the formulation. Figure A-1 shows two sample image reconstructions with respect to changing λ . While smaller λ 's provide more textured details, all λ 's achieve high-quality image reconstructions.

Besides the mixing hyperparameter λ , the sparsity hyperparameter θ choice is also key to high model performance. The importance of θ can be observed by Table A.2, in which λ is fixed to 0.7, and we observe that a non-zero θ value is always



Figure A-1: Quality of image reconstruction with various λ values

necessary for all our models and tasks to achieve the highest testing performance. Unlike λ , the range of θ values is explored on an exponential scale, and the search range heavily depends on the type of behavioral predictor. Regardless of the details, a high-dimensional latent space is often necessary for encoding urban imagery, and to enhance the generalizability of the latent space predictive power, the sparsity control through the hyperparameter θ is critical.

Table A.2: Effects of the sparsity hyperparameter θ in Model 4

<i>Panel 1: Aggregate Mode Choice - Linear Regression</i>					
θ	$1e^{-4}$	$1e^{-3}$	$1e^{-2}$	$1e^{-1}$	$1e^0$
Auto (R^2)	0.996/0.183	0.894/0.532	0.674/0.640	0.606/0.594	0.222/0.217
Active (R^2)	0.826/0.425	0.592/0.521	0.566/0.524	0.553/0.520	0.543/0.515
PT (R^2)	0.716/0.445	0.538/0.483	0.520/0.483	0.510/0.482	0.503/0.480
<i>Panel 2: Aggregate Mode Choice - Multinomial Regression ($\lambda^* = 1e^{+3}$)</i>					
θ	$1e^{+1}$	$1e^{+2}$	$1e^{+3}$	$1e^{+4}$	$1e^{+5}$
KL Loss	0.099/0.141	0.103/0.140	0.108/0.134	0.120/0.129	0.129/0.133
Auto (R^2)	0.748/0.591	0.729/0.597	0.713/0.620	0.669/0.637	0.639/0.620
Active (R^2)	0.663/0.463	0.661/0.471	0.633/0.501	0.566/0.515	0.530/0.505
PT (R^2)	0.705/0.411	0.631/0.394	0.629/0.451	0.543/0.496	0.502/0.489
<i>Panel 3: Disaggregate Mode Choice ($\lambda^* = 1e^0$)</i>					
θ	$1e^{-5}$	$1e^{-4}$	$1e^{-3}$	$1e^{-2}$	$1e^{-1}$
CE Loss	0.389/0.415	0.397/0.420	0.373/0.404	0.376/0.413	0.395/0.418
Accuracy	0.864/0.854	0.861/0.853	0.871/0.856	0.870/0.855	0.861/0.852

A.3 Two-directional imagery regeneration



Figure A-2: Another example of two-directional image regeneration

Figure A-2 demonstrates another example of reconstructing satellite images with the 2D directional movements in the latent space. Similar to Figure 3-13, we choose three targeting images, highlighted by the red squares, to generate new urban images by linear interpolation and extrapolation in the latent space. The images appear realistic, and through our framework, they are also associated with socioeconomic and behavioral information.

Appendix B

Denoising Diffusion Probabilistic Models and Latent Diffusion Models

B.1 Denoising Diffusion Probabilistic Models

Forward Process (Noising)

The forward process is defined as a Markov chain that gradually corrupts the data x_0 by adding Gaussian noise over a series of steps t from 0 to T :

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbf{I}) \quad (\text{B.1})$$

where:

- x_t is the noisy data at time step t ,
- β_t is the noise schedule at step t , a small positive value that increases slightly at each step,
- \mathcal{N} denotes the Gaussian distribution.

Given the original data x_0 , after applying the forward process for t steps, the

distribution of x_t can be expressed as:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}) \quad (\text{B.2})$$

where:

- $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$ is the cumulative product of the noise schedule.

Reverse Process (Denoising)

The reverse process is a generative model that learns to invert the forward process. It is trained to approximate the reverse conditional probabilities $p_\theta(x_{t-1}|x_t)$, which is a function of both x_t and t . A neural network, parameterized by θ , is trained to learn the parameters of the Gaussian distribution $\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)$.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (\text{B.3})$$

Training Objective

Training is performed by optimizing the usual variational lower bound on the negative log likelihood:

$$\mathbb{E} [-\log p_\theta(x_0)] \leq \mathbb{E}_q \left[-\log \frac{p_\theta(x_{0:T})}{q(x_{0:T}|x_0)} \right] \quad (\text{B.4})$$

$$= \mathbb{E}_q \left[-\log p_\theta(x_T) - \sum_{t=1}^T \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1}, x_0)} \right] \quad (\text{B.5})$$

$$(\text{B.6})$$

This loss can be further decomposed as

$$\mathbb{E}_q \left[\underbrace{D_{KL}(q(x_T|x_0)||p(x_T))}_{L_T} + \sum_{t>1} \underbrace{D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))}_{L_{t-1}} - \underbrace{\log p_\theta(x_0|x_1)}_{L_0} \right] \quad (\text{B.7})$$

In particular, the term L_{t-1} trains the network to perform on one reverse diffusion step and is denoted as L_{vlb} . This KL divergence is tractable since p and q are both Gaussian. However, compared to directly predicting $\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)$, it is easier to predict the cumulative noise $\epsilon_\theta(x_t, t)$ added to the intermediate x_t [59]. Thus the predicted mean $\mu_\theta(x_t, t)$ is parameterized as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{1 - \beta_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) \quad (\text{B.8})$$

Additionally, [59] found that fixing $\Sigma_\theta(x_t, t)$ to the noise schedule β_t does not affect the generated sample quality [59]. Under this parametrization, the loss term can be simplified to

$$L_{simple}(\theta) = \mathbb{E}_{t,x,\epsilon} [\|\epsilon - \epsilon_\theta(x_t, t)\|^2] \quad (\text{B.9})$$

where λ balances the learning of the mean and the variance.

B.2 Latent Diffusion Models (Stable Diffusion)

Since diffusion models operate directly in the pixel space with thousands of diffusion steps, they are very resource-intensive to train. Latent diffusion models, commonly known as stable diffusion, operate in the latent space z of a powerful pre-trained autoencoder. The rest is the same and the training objective now reads

$$L_{LDM}(\theta) = \mathbb{E}_{t,\mathcal{E}(x),\epsilon} [\|\epsilon - \epsilon_\theta(z_t, t)\|^2] \quad (\text{B.10})$$

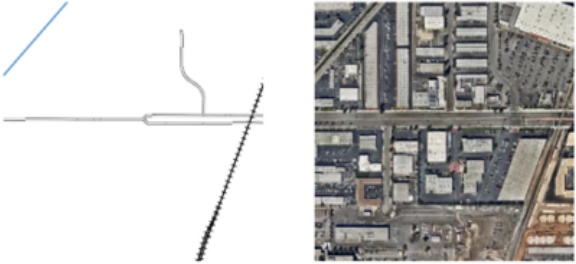
where \mathcal{E} is the encoder. At inference time, samples from $p(z)$ can be restored to the image space by passing them through the decoder \mathcal{D} .

Appendix C

Generative Design - User Study

This appendix presents a sample of the questions for Parts 1 (Figure C-1) and 2 (Figure C-2) of the user study.

Prompt: Satellite image of a town in Los Angeles. Landuse parcels include 5 percent residential, 15 percent industrial, 80 percent commercial, 5 percent open parking. Residential area consists entirely of houses. Dense building coverage.



Rate the land use patterns. *

	1	2	3	4	5	
Incompatible	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Compatible

Rate the adherence to constraints presented. *

	1	2	3	4	5	
Violated	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Well-respected

Rate the realism of images. *

	1	2	3	4	5	
Very fake	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very realistic

Figure C-1: Generative Urban Design - User Study Part 1

Prompt: Satellite image of a city in Los Angeles. Landuse parcels include 20 percent residential, 45 percent commercial, 15 percent park, 10 percent open parking. Residential area consists entirely of houses. Medium building coverage.



Design 1



Design 2

Which aerial image reveals an urban environment that is closer to the language description? *

- Design 1
- Design 2
- Both are good.
- Both are bad.

Figure C-2: Generative Urban Design - User Study Part 2

Bibliography

- [1] My Daily Travel Survey, 2018-2019: Public Data - Datasets - CMAP Data Hub.
- [2] Adrian Albert, Jasleen Kaur, and Marta C González. Using Convolutional Networks and Satellite Imagery to Identify Patterns in Urban Environments at a Large Scale. 2017.
- [3] Melissa R. Allen-Dumas, Abigail R. Wheelis, Levi T. Sweet-Breu, Joshua Anantharaj, and Kuldeep R. Kurte. Generative adversarial networks for ensemble projections of future urban morphology. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities*, page 1–6, Seattle Washington, 2022. ACM.
- [4] Ahmad Alwosheel, Sander van Cranenburgh, and Caspar G Chorus. Why did you predict that? towards explainable artificial neural networks for travel demand analysis. *Transportation Research Part C: Emerging Technologies*, 128:103143, 2021.
- [5] Ioanna Arkoudi, Carlos Lima Azevedo, and Francisco C. Pereira. Combining discrete choice models and neural networks through embeddings: Formulation, interpretability and performance. 9 2021.
- [6] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics*, 42(4):149:1–149:11, July 2023.
- [7] Kumar Ayush, Burak Uzkent, Marshall Burke, David Lobell, and Stefano Ermon. Generating interpretable poverty maps using object detection in satellite images. *IJCAI International Joint Conference on Artificial Intelligence*, 2021-Janua:4410–4416, 2020.
- [8] Poopak Azad, Nima Jafari Navimipour, Amir Masoud Rahmani, and Arash Sharifi. The role of structured and unstructured data managing mechanisms in the internet of things. *Cluster Computing*, 23(2):1185–1198, 2020.
- [9] Prateek Bansal, Kara M Kockelman, and Amit Singh. Assessing public opinions of and interest in new vehicle technologies: An austin perspective. *Transportation Research Part C: Emerging Technologies*, 67:1–14, 2016.

- [10] Michael Batty, David Chapman, Steve Evans, Mordechai Haklay, Stefan Kueppers, Naru Shiode, Andy Smith, and Paul Torrens. Visualising the city: Communicating urban design to planners and decision-makers. October 2000.
- [11] Moshe Ben-Akiva, Daniel Mcfadden, Kenneth Train, Joan Walker, Chandra Bhat, Michel Bierlaire, Ecole Polytechnique, Fédérale De Lausanne, Axel Boersch-Supan, David Brownstone, David S Bunch, Andrew Daly, Rand Europe, Andre De Palma, and Marcela A Munizaga. Hybrid Choice Models: Progress and Challenges. *Marketing Letters*, 13:163–175, 2002.
- [12] Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: a methodological tour d’horizon. *European Journal of Operational Research*, 290(2):405–421, 2021.
- [13] Angela Stefania Bergantino, Mauro Capurso, and Stephane Hess. Modelling regional accessibility to airports using discrete choice models: An application to a system of regional airports. *Transportation Research Part A: Policy and Practice*, 132:855–871, 2020.
- [14] David Berthelot, Ian Goodfellow, Colin Raffel, and Aurko Roy. Understanding and improving interpolation in autoencoders via an adversarial regularizer. *7th International Conference on Learning Representations, ICLR 2019*, pages 1–20, 2019.
- [15] Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.
- [16] Filip Biljecki and Koichi Ito. Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning*, 215:104217, 11 2021.
- [17] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *32nd International Conference on Machine Learning, ICML 2015*, 2:1613–1622, 2015.
- [18] US Census Bureau. Urban and rural.
- [19] Marshall Burke, Anne Driscoll, David B. Lobell, and Stefano Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, March 2021.
- [20] Giulio Erberto Cantarella and Stefano de Luca. Multilayer feedforward networks for transportation mode choice analysis: An analysis and a comparison with random utility models. *Transportation Research Part C: Emerging Technologies*, 13(2):121–155, 2005. Handling Uncertainty in the Analysis of Traffic and Transportation Systems (Bari, Italy, June 10–13 2002).

- [21] Dun Cao, Kai Zeng, Jin Wang, Pradip Kumar Sharma, Xiaomin Ma, Yonghe Liu, and Siyuan Zhou. Bert-based deep spatial-temporal network for taxi demand prediction. *IEEE Transactions on Intelligent Transportation Systems*, 2021.
- [22] Denise Capasso Da Silva, David A. King, and Shea Lemar. Accessibility in practice: 20-minute city as a sustainability planning goal. *Sustainability*, 12(1):129, January 2020.
- [23] Matthew Carmona. *Public Places Urban Spaces: The Dimensions of Urban Design*. Routledge, 3 edition, February 2021.
- [24] R. Caruana and V. D. Sa. Promoting Poor Features to Supervisors: Some Inputs Work Better as Outputs. *undefined*, 1996.
- [25] Enhui Chen, Zhirui Ye, Chao Wang, and Mingtao Xu. Subway Passenger Flow Prediction for Special Events Using Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):1109–1120, mar 2020.
- [26] Yulu Chen, Rongjun Qin, Guixiang Zhang, and Hessah Albanwan. Spatial temporal analysis of traffic patterns during the covid-19 epidemic by vehicle detection using planet remote-sensing satellite images. *Remote Sensing*, 13(2):1–18, 1 2021.
- [27] Mario Cools, Bruno Kochan, Tom Bellemans, Davy Janssens, and Geert Wets. Assessment of the effect of micro-simulation error on key travel indices: evidence from the activity-based model feathers. *90th Annual Meeting of the Transportation Research Board*, pages 1–15, 2010.
- [28] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, September 2023.
- [29] Mercy Dada, Mark Zuidgeest, and Stephane Hess. Modelling pedestrian crossing choice on cape town’s freeways: Caught between a rock and a hard place? *Transportation research part F: traffic psychology and behaviour*, 60:245–261, 2019.
- [30] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, page 8780–8794. Curran Associates, Inc., 2021.
- [31] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.
- [32] Weishan Dong, Ting Yuan, Kai Yang, Changsheng Li, and Shilei Zhang. Autoencoder regularized network for driving style representation learning. *arXiv preprint arXiv:1701.01272*, 2017.

- [33] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In *European conference on computer vision*, pages 196–212. Springer, 2016.
- [34] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017.
- [35] Yuki Endo, Hiroyuki Toda, Kyosuke Nishida, and Akihisa Kawanobe. Deep feature extraction from trajectories for transportation mode estimation. In *Advances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part II 20*, pages 54–66. Springer, 2016.
- [36] Miguel Espinosa and Elliot J. Crowley. Generate your own scotland: Satellite image generation conditioned on maps. (arXiv:2308.16648), August 2023. arXiv:2308.16648 [cs].
- [37] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, June 2021.
- [38] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. *Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors*, volume 13675, page 89–106. Springer Nature Switzerland, Cham, 2022.
- [39] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *33rd International Conference on Machine Learning, ICML 2016*, 2016.
- [40] Xiaowei Gao, Xinke Jiang, Dingyi Zhuang, Huanfa Chen, Shenhao Wang, and James Haworth. Spatiotemporal graph neural networks with uncertainty quantification for traffic incident risk prediction. *arXiv preprint arXiv:2309.05072*, 2023.
- [41] Jakob Gawlikowski, Cedric Robile Njjeutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A Survey of Uncertainty in Deep Neural Networks. jul 2021.
- [42] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the United

- States. *Proceedings of the National Academy of Sciences of the United States of America*, 114(50):13108–13113, 2017.
- [43] Xu Geng, Yaguang Li, Leye Wang, Lingyu Zhang, Qiang Yang, Jieping Ye, and Yan Liu. Spatiotemporal Multi-Graph Convolution Network for Ride-Hailing Demand Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:3656–3663, 2019.
- [44] Aurélie Glerum, Lidija Stankovikj, Michaël Thémans, and Michel Bierlaire. Forecasting the demand for electric vehicles: Accounting for attitudes and perceptions. <https://doi.org/10.1287/trsc.2013.0487>, 48:483–499, 12 2013. HCM application.
- [45] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [46] Helmut Grabner, Peter M Roth, and Horst Bischof. Eigenboosting: Combining discriminative and generative information. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.
- [47] Jianhua Guo, Wei Huang, and Billy M. Williams. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. *Transportation Research Part C: Emerging Technologies*, 43:50–64, 2014.
- [48] Xiaotong Guo, Nicholas S. Caros, and Jinhua Zhao. Robust matching-integrated vehicle rebalancing in ride-hailing system with uncertain demand. *Transportation Research Part B: Methodological*, 150:161–189, aug 2021.
- [49] Xiaotong Guo, Qingyi Wang, and Jinhua Zhao. Data-driven vehicle rebalancing with predictive prescriptions in the ride-hailing system. *IEEE Open Journal of Intelligent Transportation Systems*, 3:251–266, 2022.
- [50] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv:1704.03477*, 2017.
- [51] Yafei Han, Christopher Zegras, Francisco Camara Pereira, and Moshe Ben-Akiva. A neural-embedded choice model: Tastenet-mnl modeling taste heterogeneity with flexibility and interpretability. *arXiv preprint arXiv:2002.00922*, 2020.
- [52] Steve Hankey, Wenwen Zhang, Huyen TK Le, Perry Hystad, and Peter James. Predicting bicycling and walking traffic using street view imagery and destination data. *Transportation research part D: transport and environment*, 90:102651, 2021.

- [53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.
- [54] John Paul Helveston, Yimin Liu, Elea McDonnell Feit, Erica Fuchs, Erica Klampfl, and Jeremy J Michalek. Will subsidies drive electric vehicle adoption? measuring consumer preferences in the us and china. *Transportation Research Part A: Policy and Practice*, 73:96–112, 2015.
- [55] Tom Heskes. Practical confidence and prediction intervals. In *Advances in Neural Information Processing Systems*, 1997.
- [56] Stephane Hess, Greg Spitz, Mark Bradley, and Matt Coogan. Analysis of mode choice for intercity travel: Application of a hybrid choice model to two distinct us corridors. *Transportation Research Part A: Policy and Practice*, 116:547–567, 2018.
- [57] Tim Hillel. New perspectives on the performance of machine learning classifiers for mode choice prediction. Technical report, 2020.
- [58] Tim Hillel, Michel Bierlaire, Mohammed Z.E.B. Elshafie, and Ying Jin. A systematic review of machine learning classification methodologies for modelling passenger mode choice. *Journal of Choice Modelling*, 38:100221, 3 2021.
- [59] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, page 6840–6851. Curran Associates, Inc., 2020.
- [60] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [61] Ning Huan, Stephane Hess, and Enjian Yao. Understanding the effects of travel demand management on metro commuters’ behavioural loyalty: a hybrid choice modelling approach. *Transportation*, pages 1–30, 2 2021.
- [62] Ning Huan, Stephane Hess, and Enjian Yao. Understanding the effects of travel demand management on metro commuters’ behavioural loyalty: a hybrid choice modelling approach. *Transportation*, 49(2):343–372, 2022.
- [63] Ming Huang, Fuzhen Zhuang, Xiao Zhang, Xiang Ao, Zhengyu Niu, Min-Ling Zhang, and Qing He. Supervised representation learning for multi-label classification. *Machine Learning*, 108:747–763, 2019.
- [64] Rowan T. Hughes, Liming Zhu, and Tomasz Bednarz. Generative adversarial networks-enabled human-artificial intelligence collaborative applications for creative and design industries: A systematic review of current approaches and trends. *Frontiers in Artificial Intelligence*, 4, 2021.

- [65] Mohamed R. Ibrahim, James Haworth, and Tao Cheng. Understanding cities with machine eyes: A review of deep computer vision in urban analytics. *Cities*, 96:102481, 1 2020.
- [66] Muhammad Zudhy Irawan, Prawira Fajarindra Belgiawan, Tri Basuki Joewono, and Nurvita I.M. Simanjuntak. Do motorcycle-based ride-hailing apps threaten bus ridership? a hybrid choice modeling approach with latent variables. *Public Transport*, 12:207–231, 3 2020.
- [67] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.
- [68] Feifeng Jiang, Jun Ma, Christopher John Webster, Alain J.F. Chiaradia, Yulun Zhou, Zhan Zhao, and Xiaohu Zhang. Generative urban design: A systematic review on problem formulation, design generation, and decision-making. *Progress in Planning*, page 100795, 2023.
- [69] Xinke Jiang, Dingyi Zhuang, Xianghui Zhang, Hao Chen, Jiayuan Luo, and Xiaowei Gao. Uncertainty quantification via spatial-temporal tweedie model for zero-inflated and long-tail travel demand prediction. *arXiv preprint arXiv:2306.09882*, 2023.
- [70] H. M.Dipu Kabir, Abbas Khosravi, Mohammad Anwar Hosen, and Saeid Nahavandi. Neural Network-Based Uncertainty Quantification: A Survey of Methodologies and Applications. *IEEE Access*, 6:36218–36234, 2018.
- [71] M. G. Karlaftis and E. I. Vlahogianni. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19:387–399, 2011. comprehensive comparison of NN and statistical methods both in terms of philosophy and transportation applications.
- [72] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. page 8110–8119, 2020.
- [73] Jintao Ke, Hongyu Zheng, Hai Yang, and Xiqun Chen. Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. *Transportation Research Part C-emerging Technologies*, 2017.
- [74] Annet Kempenaar, Judith Westerink, Marjo van Lierop, Marlies Brinkhuijsen, and Adri van den Brink. “design makes you understand”—mapping the contributions of designing to regional planning and development. *Landscape and Urban Planning*, 149:20–30, May 2016.

- [75] Abbas Khosravi and Saeid Nahavandi. An optimized mean variance estimation method for uncertainty quantification of wind power forecasts. *International Journal of Electrical Power & Energy Systems*, 61:446–454, 2014.
- [76] Abbas Khosravi, Saeid Nahavandi, Doug Creighton, and Amir F. Atiya. Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE Transactions on Neural Networks*, 22(3):337–346, 2011.
- [77] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pages 1–14, 2017.
- [78] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112, 2009.
- [79] Danyel Koca, Jan Dirk Schmöcker, and Kouji Fukuda. Origin-destination matrix estimation by deep learning using maps with new york case study. In *2021 7th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 1–6. IEEE, 2021.
- [80] Roger Koenker and Kevin F. Hallock. Quantile regression. *Journal of Economic Perspectives*, 2001.
- [81] James Kotary, Ferdinando Fioretto, Pascal Van Hentenryck, and Bryan Wilder. End-to-end constrained optimization learning: A survey. *arXiv preprint arXiv:2103.16378*, 2021.
- [82] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):7386–7400, 2021.
- [83] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 43(2):145–150, dec 2016.
- [84] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016.
- [85] Julia A. Lasserre, Christopher M. Bishop, and Thomas P. Minka. Principled hybrids of generative and discriminative models. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1(6):87–94, 2006.
- [86] Lei Le, Andrew Patterson, and Martha White. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS):107–117, 2018.

- [87] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.
- [88] Dongwoo Lee, Sybil Derrible, and Francisco Camara Pereira. Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling. *Transportation Research Record*, 2672(49):101–112, 2018.
- [89] Can Li, Lei Bai, Wei Liu, Lina Yao, and S. Travis Waller. Graph Neural Network for Robust Public Transit Demand Prediction. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–13, 2020.
- [90] Hongzhou Lin, Joshua Robinson, and Stefanie Jegelka. Perceptual regularization: Visualizing and learning generalizable representations. 2019.
- [91] Fuqiang Liu, Jiawei Wang, Jingbo Tian, Dingyi Zhuang, Luis Miranda-Moreno, and Lijun Sun. A universal framework of spatiotemporal bias block for long-term traffic forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 2022.
- [92] Lingbo Liu, Jingwen Chen, Hefeng Wu, Jiajie Zhen, Guanbin Li, and Liang Lin. Physical-Virtual Collaboration Modeling for Intra- and Inter-Station Metro Ridership Prediction. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–15, 2020.
- [93] Yang Liu, Zhiyuan Liu, Cheng Lyu, and Jieping Ye. Attention-based deep ensemble net for large-scale online taxi-hailing demand prediction. *IEEE transactions on intelligent transportation systems*, 21(11):4798–4807, 2019.
- [94] Yang Liu, Cheng Lyu, Yuan Zhang, Zhiyuan Liu, Wenwu Yu, and Xiaobo Qu. Deeptsp: Deep traffic state prediction model based on large-scale empirical data. *Communications in transportation research*, 1:100012, 2021.
- [95] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. page 11461–11471, 2022.
- [96] Kevin Lynch. *The image of the city*. Publication of the Joint Center for Urban studies. M.I.T. Press, Cambridge, Mass., 33. print edition, 2008.
- [97] Alireza Mahpour, Amirreza Mamdoohi, Taha HosseinRashidi, Basil Schmid, and Kay W Axhausen. Shopping destination choice in tehran: An integrated choice and latent variable approach. *Transportation research part F: traffic psychology and behaviour*, 58:566–580, 2018.
- [98] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

- [99] Hafsa Maryam, Tania Panayiotou, and Georgios Ellinas. Uncertainty quantification and consideration in ml-aided traffic-driven service provisioning. *Computer Communications*, 202:13–22, 2023.
- [100] Andreas Maurer, Bernardino Romera-Paredes, Urun Dogan, Marius Kloft, Francesco Orabona, and Tatiana Tommasi. The benefit of multitask representation learning. *jmlr.org*, 17:1–32, 2016.
- [101] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. 1974.
- [102] Daniel McFadden. Modeling the choice of residential location. *Transportation Research Record*, (673), 1978.
- [103] Daniel McFadden and Kenneth Train. Mixed mnl models for discrete response. *Journal of applied Econometrics*, 15(5):447–470, 2000.
- [104] Larry R Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5(64-67):2, 2001.
- [105] Hugues Moreau, Andrea Vassilev, and Liming Chen. Data fusion for deep learning on transport mode detection: A case study. In *International Conference on Engineering Applications of Neural Networks*, pages 141–152. Springer, 2021.
- [106] Carlos Moreno, Zaheer Allam, Didier Chabaud, Catherine Gall, and Florent Pratlong. Introducing the “15-minute city”: Sustainability, resilience and place identity in future post-pandemic cities. *Smart Cities*, 4(1):93–111, March 2021.
- [107] Johannes Mueller, Hangxin Lu, Artem Chirkin, Bernhard Klein, and Gerhard Schmitt. Citizen design science: A strategy for crowd-creative urban design. *Cities*, 72:181–188, 2018.
- [108] Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L. Glaeser, and César A. Hidalgo. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences of the United States of America*, 114(29):7571–7576, 2017.
- [109] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 779–785, 2014.
- [110] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, page 8162–8171. PMLR, July 2021.
- [111] David A. Nix and Andreas S. Weigend. Estimating the mean and variance of the target probability distribution. In *IEEE International Conference on Neural Networks - Conference Proceedings*, 1994.

- [112] Ariel Noyman and Kent Larson. A deep image of the city: generative urban-design visualization. In *Proceedings of the 11th Annual Symposium on Simulation for Architecture and Urban Design*, SimAUD '20, page 1–8, San Diego, CA, USA, May 2020. Society for Computer Simulation International.
- [113] Alon Oring, Zohar Yakhini, and Yacov Hel-Or. Autoencoder image interpolation by shaping the latent space. *arXiv preprint arXiv:2008.01487*, 2020.
- [114] Miguel Paredes, Erik Hemberg, Una-May O'Reilly, and Chris Zegras. Machine learning or discrete choice models for car ownership demand estimation and prediction? In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 780–785. IEEE, 2017.
- [115] Chulwoong Park, Wonjun No, Junyong Choi, and Youngchul Kim. Development of an ai advisor for conceptual land use planning. *Cities*, 138:104371, 2023.
- [116] Axel Parmentier. Learning to approximate industrial problems by operations research classic problems. *Operations Research*, 70(1):606–623, 2022.
- [117] Tim Pearce, Mohamed Zaki, Alexandra Brintrup, and Andy Neely. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. *35th International Conference on Machine Learning, ICML 2018*, 9:6473–6482, 2018.
- [118] Olga Petrik, Filipe Moura, and João de Abreu e. Silva. Measuring uncertainty in discrete choice travel demand forecasting models. *Transportation Planning and Technology*, 39(2):218–237, feb 2016.
- [119] Walter H. L. Pinaya, Mark S. Graham, Robert Gray, Pedro F. da Costa, Petru-Daniel Tudosiu, Paul Wright, Yee H. Mah, Andrew D. MacKinnon, James T. Teo, Rolf Jager, David Werring, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Fast unsupervised brain anomaly detection and segmentation with diffusion models. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Lecture Notes in Computer Science, page 705–714, Cham, 2022. Springer Nature Switzerland.
- [120] Weizhu Qian, Dalin Zhang, Yan Zhao, Kai Zheng, and JQ James. Uncertainty quantification for traffic forecasting: A unified approach. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 992–1004. IEEE, 2023.
- [121] Zhiyu Qu, Tao Xiang, and Yi-Zhe Song. Sketchdreamer: Interactive text-augmented creative sketch ideation. (arXiv:2308.14191), August 2023. arXiv:2308.14191 [cs].

- [122] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, page 8748–8763. PMLR, July 2021.
- [123] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. (arXiv:2204.06125), April 2022. arXiv:2204.06125 [cs].
- [124] Haziq Razali, Taylor Mordan, and Alexandre Alahi. Pedestrian intention prediction: A convolutional bottom-up multi-task approach. *Transportation research part C: emerging technologies*, 130:103259, 2021.
- [125] Lei Ren, Yaqiang Sun, Jin Cui, and Lin Zhang. Bearing remaining useful life prediction based on deep autoencoder and deep neural networks. *Journal of Manufacturing Systems*, 48:71–77, 2018.
- [126] Filipe Rodrigues and Francisco C. Pereira. Beyond Expectation: Deep Joint Mean and Quantile Regression for Spatiotemporal Problems. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13, 2020.
- [127] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*, 12(1):4392, July 2021.
- [128] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.
- [129] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. page 10684–10695, 2022.
- [130] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, December 2022.
- [131] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, April 2023.

- [132] Divya Saxena and Jiannong Cao. Generative adversarial networks (gans): Challenges, solutions, and future directions. *ACM Computing Surveys*, 54(3):63:1–63:42, May 2021.
- [133] Basil Schmid and Kay W. Axhausen. In-store or online shopping of search and experience goods: A hybrid choice approach. *Journal of Choice Modelling*, 31:156–180, 6 2019.
- [134] Ian Seiferling, Nikhil Naik, Carlo Ratti, and Raphaël Proulx. Green streets - Quantifying and mapping urban trees with street-level imagery and computer vision. *Landscape and Urban Planning*, 165(July 2016):93–101, 2017.
- [135] Hongzhi Shi, Quanming Yao, Qi Guo, Yaguang Li, Lingyu Zhang, Jieping Ye, Yong Li, and Yan Liu. Predicting origin-destination flow via multi-perspective graph convolutional network. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1818–1821. IEEE, 2020.
- [136] Joachim Sicking, Maram Akila, Maximilian Pintz, Tim Wirtz, Asja Fischer, and Stefan Wrobel. A Novel Regression Loss for Non-Parametric Uncertainty Optimization. pages 2021–2022, jan 2021.
- [137] Brian Sifringer, Virginie Lurkin, and Alexandre Alahi. Enhancing discrete choice models with representation learning. *Transportation Research Part B: Methodological*, 140:236–261, 10 2020.
- [138] Kenneth A Small and Harvey S Rosen. Applied welfare economics with discrete choice models. *Econometrica: Journal of the Econometric Society*, pages 105–130, 1981.
- [139] Jose J Soto, Victor Cantillo, and Julian Arellana. Incentivizing alternative fuel vehicles: the influence of transport policies, attitudes and perceptions. *Transportation*, 45:1721–1753, 2018.
- [140] Kenneth Train. A structured logit model of auto ownership and mode choice. *The Review of Economic Studies*, 47(2):357–370, 1980.
- [141] Sander van Cranenburgh and Ahmad Alwosheel. An artificial neural network based approach to investigate travellers’ decision rules. *Transportation Research Part C: Emerging Technologies*, 98:152–166, 2019.
- [142] Sander van Cranenburgh and Marco Kouwenhoven. An artificial neural network based method to uncover the value-of-travel-time distribution. *Transportation*, 48(5):2545–2583, 2021.
- [143] Sander van Cranenburgh, Shenhao Wang, Akshay Vij, Francisco Pereira, and Joan Walker. Choice modelling in the age of machine learning - discussion paper. *Journal of Choice Modelling*, 42:100340, 3 2022.

- [144] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [145] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- [146] Akshay Vij and Joan L. Walker. How, when and why integrated choice and latent variable models are latently useful. *Transportation Research Part B: Methodological*, 90:192–217, 8 2016.
- [147] Eleni Vlahogianni and Matthew Karlaftis. Temporal aggregation in traffic data: implications for statistical characteristics and model choice. *Transportation Letters*, 3(1):37–49, 2011.
- [148] Joan Walker and Moshe Ben-Akiva. Generalized random utility model. *Mathematical social sciences*, 43:303–343, 2002.
- [149] Dongjie Wang, Yanjie Fu, Kunpeng Liu, Fanglan Chen, Pengyang Wang, and Chang-Tien Lu. Automated urban planning for reimagining city configuration via adversarial learning: Quantification, generation, and evaluation. *ACM Transactions on Spatial Algorithms and Systems*, 9(1):1–24, March 2023.
- [150] Dongjie Wang, Lingfei Wu, Denghui Zhang, Jingbo Zhou, Leilei Sun, and Yanjie Fu. Human-instructed deep hierarchical generative learning for automated urban planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4660–4667, June 2023.
- [151] Jing Wang, Shengnan Guo, Tonglong Wei, Yiji Zhao, Youfang Lin, Huaiyu Wan, et al. Spatial-temporal uncertainty-aware graph networks for promoting accuracy and reliability of traffic forecasting.
- [152] Shenhao Wang, Baichuan Mo, and Jinhua Zhao. Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions. *Transportation Research Part C: Emerging Technologies*, 112:234–251, 2020.
- [153] Shenhao Wang, Baichuan Mo, and Jinhua Zhao. Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions. *Transportation Research Part C: Emerging Technologies*, 112:234–251, 3 2020.
- [154] Shenhao Wang, Baichuan Mo, and Jinhua Zhao. Theory-based residual neural networks: A synergy of discrete choice models and deep neural networks. *Transportation Research Part B: Methodological*, 146:333–358, 4 2021.
- [155] Shenhao Wang, Qingyi Wang, Nate Bailey, and Jinhua Zhao. Deep neural networks for choice analysis: A statistical learning theory perspective. *Transportation Research Part B: Methodological*, 148:60–81, 2021.

- [156] Shenhao Wang, Qingyi Wang, and Jinhua Zhao. Deep neural networks for choice analysis: Extracting complete economic information for interpretation. *Transportation Research Part C: Emerging Technologies*, 118:102701, 2020.
- [157] Shenhao Wang, Qingyi Wang, and Jinhua Zhao. Deep neural networks for choice analysis: Extracting complete economic information for interpretation. *Transportation Research Part C: Emerging Technologies*, 118:102701, 2020.
- [158] Shenhao Wang, Qingyi Wang, and Jinhua Zhao. Multitask learning deep neural networks to combine revealed and stated preference data. *Journal of Choice Modelling*, page 100236, 2020.
- [159] Shenhao Wang and Jinhua Zhao. Risk preference and adoption of autonomous vehicles. *Transportation research part A: policy and practice*, 126:215–229, 2019.
- [160] Wei Wang and Yiwei Wu. Is uncertainty always bad for the performance of transportation systems? *Communications in transportation research*, 1:100021, 2021.
- [161] Zepu Wang, Dingyi Zhuang, Yankai Li, Jinhua Zhao, and Peng Sun. St-gin: An uncertainty quantification approach in traffic data imputation with spatio-temporal graph attention and bidirectional recurrent neural networks. *arXiv preprint arXiv:2305.06480*, 2023.
- [162] Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. (arXiv:2305.01115), October 2023. arXiv:2305.01115 [cs].
- [163] Min Weng, Ning Ding, Jing Li, Xianfeng Jin, He Xiao, Zhiming He, and Shiliang Su. The 15-minute walkable neighborhoods: Measurement, social inequalities and implications for building healthy communities in urban china. *Journal of Transport and Health*, 13:259–273, June 2019.
- [164] OpenStreetMap Wiki. Slippy map tilenames - openstreetmap wiki.
- [165] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C. Cattin. Diffusion models for implicit image segmentation ensembles. In *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, page 1336–1348. PMLR, December 2022.
- [166] Melvin Wong and Bilal Farooq. Modelling latent travel behaviour characteristics with generative machine learning. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 749–754. IEEE, 2018.
- [167] Melvin Wong and Bilal Farooq. Reslogit: A residual neural network logit model for data-driven choice modelling. *Transportation Research Part C: Emerging Technologies*, 126:103050, 2021.

- [168] Abraham Noah Wu, Rudi Stouffs, and Filip Biljecki. Generative adversarial networks in the built environment: A comprehensive review of the application of gans across data types and scales. *Building and Environment*, 223:109477, 2022.
- [169] Yuankai Wu, Dingyi Zhuang, Aurelie Labbe, and Lijun Sun. Inductive graph neural networks for spatiotemporal kriging. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4478–4485, May 2021.
- [170] Yuankai Wu, Dingyi Zhuang, Mengying Lei, Aurelie Labbe, and Lijun Sun. Spatial aggregation and temporal convolution networks for real-time kriging. *arXiv preprint arXiv:2109.12144*, 2021.
- [171] Xi Xiong, Kaan Ozbay, Li Jin, and Chen Feng. Dynamic origin–destination matrix prediction with line graph neural networks and kalman filter. *Transportation Research Record*, 2674(8):491–503, 2020.
- [172] Jiayang Xu and Karthik Duraisamy. Multi-level convolutional autoencoder networks for parametric prediction of spatio-temporal dynamics. *Computer Methods in Applied Mechanics and Engineering*, 372:113379, 2020.
- [173] Wenju Xu, Shawn Keshmiri, and Guanghui Wang. Adversarially approximated autoencoder for image generation and manipulation. *IEEE Transactions on Multimedia*, 21(9):2387–2396, 2019.
- [174] Kun Yang, Justin Tu, and Tian Chen. Homoscedasticity: An overlooked critical assumption for linear regression. *General Psychiatry*, 32(5):100148, oct 2019.
- [175] Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, Zhenhui Li, Jieping Ye, and Didi Chuxing. Deep multi-view spatial-temporal network for taxi demand prediction. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 2588–2595, 2018.
- [176] Rui Yao and Shlomo Bekhor. A variational autoencoder approach for choice set generation and implicit perception of alternatives in choice modeling. *Transportation Research Part B: Methodological*, 158:273–294, 2022.
- [177] Ali Yazdizadeh, Zachary Patterson, and Bilal Farooq. Ensemble convolutional neural networks for mode inference in smartphone travel survey. *IEEE Transactions on Intelligent Transportation Systems*, 21(6):2232–2239, 2019.
- [178] Ali Yazdizadeh, Zachary Patterson, and Bilal Farooq. Semi-supervised gans to infer travel modes in gps trajectories. *Journal of Big Data Analytics in Transportation*, 3:201–211, 2021.
- [179] Jiexia Ye, Juanjuan Zhao, Kejiang Ye, and Chengzhong Xu. How to Build a Graph-Based Deep Learning Architecture in Traffic Domain: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

- [180] Jiexia Ye, Juanjuan Zhao, Kejiang Ye, and Chengzhong Xu. Multi-STGCnet: A Graph Convolution Based Spatial-Temporal Framework for Subway Passenger Flow Forecasting. In *Proceedings of the International Joint Conference on Neural Networks*. Institute of Electrical and Electronics Engineers Inc., jul 2020.
- [181] Xinyue Ye, Jiaxin Du, and Yu Ye. Masterplangan: Facilitating the smart rendering of urban master plans via generative adversarial networks. *Environment and Planning B: Urban Analytics and City Science*, 49(3):794–814, 2022.
- [182] Yang Ye and Shihao Ji. Sparse graph attention networks. *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [183] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11(1):1–11, 2020.
- [184] Luca Zamparini and Aura Reggiani. *The value of travel time in passenger and freight transport: an overview*, pages 161–178. Routledge, 2016.
- [185] Fan Zhang, Bolei Zhou, Liu Liu, Yu Liu, Helene H Fung, Hui Lin, and Carlo Ratti. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, 180:148–160, 2018.
- [186] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. page 3836–3847, 2023.
- [187] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society.
- [188] Yanru Zhang, Ranye Sun, Ali Haghani, and Xiaosi Zeng. Univariate volatility-based models for improving quality of travel time reliability forecasting. *Transportation research record*, 2365(1):73–81, 2013.
- [189] Yong Zhao and Kara Maria Kockelman. The propagation of uncertainty through travel demand models: An exploratory analysis. *Annals of Regional Science*, 36(1):145–163, 2002.
- [190] Yu Zheng, Yuming Lin, Liang Zhao, Tinghai Wu, Depeng Jin, and Yong Li. Spatial planning of urban communities via deep reinforcement learning. *Nature Computational Science*, 3(9):748–762, September 2023.
- [191] Yunhan Zheng, Shenhao Wang, and Jinhua Zhao. Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models. *Transportation Research Part C: Emerging Technologies*, 132:103410, 2021.

- [192] Dingyi Zhuang, Yuheng Bu, Guang Wang, Shenhao Wang, and Jinhua Zhao. SAUC: Sparsity-aware uncertainty calibration for spatiotemporal prediction with graph neural networks. In *Temporal Graph Learning Workshop @ NeurIPS 2023*, 2023.
- [193] Dingyi Zhuang, Siyu Hao, Der-Horng Lee, and Jian Gang Jin. From compound word to metropolitan station: Semantic similarity analysis using smart card data. *Transportation Research Part C: Emerging Technologies*, 114:322–337, 2020.
- [194] Dingyi Zhuang, Shenhao Wang, Haris Koutsopoulos, and Jinhua Zhao. Uncertainty quantification of sparse travel demand prediction with spatial-temporal graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4639–4647, 2022.
- [195] Roland S. Zimmermann, Lukas Schott, Yang Song, Benjamin A. Dunn, and David A. Klindt. Score-based generative classifiers. (arXiv:2110.00473), December 2021. arXiv:2110.00473 [cs, stat].