

Article

Generating Large-Scale Origin–Destination Matrix via Progressive Growing Generative Adversarial Networks Model

Zehao Yuan ^{1,2,3,*}, Xuanyan Chen ^{1,2,3}, Biyu Chen ^{1,2,3} , Yubo Luo ⁴, Yu Zhang ^{1,2,3}, Wenxin Teng ^{1,2,3} and Chao Zhang ^{1,2,3}

- ¹ State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; xuanyanc@whu.edu.cn (X.C.); chen.biyu@whu.edu.cn (B.C.); yu.zhang@whu.edu.cn (Y.Z.); tengwenxin1112@126.com (W.T.); zhangchao1@whu.edu.cn (C.Z.)
² Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China
³ Geocomputation Center for Social Sciences, Wuhan University, Wuhan 430079, China
⁴ Hangzhou Institute of Technology, Xidian University, Hangzhou 311200, China; luoyubo@xidian.edu.cn
* Correspondence: yuanzehao@whu.edu.cn

Abstract: The origin–destination (OD) matrix describes traffic flow information between regions. It is a critical input for intelligent transportation systems (ITS). However, obtaining the OD matrix remains challenging due to high costs and privacy concerns. Synthetic data, which have the same statistical distribution of real data, help address privacy issues and data scarcity. Based on Generative Adversarial Networks (GAN), OD matrix generation models, which can effectively generate a synthetic OD matrix, help to address the challenge of obtaining OD matrix data in ITS research. However, existing OD matrix generation methods can only handle with tens of nodes. To address this challenge, this study proposes the Origin–Destination Progressive Growing Generative Adversarial Networks (OD-PGGAN) for large-scale OD matrix generation task which adapt the PGGAN architecture. OD-PGGAN adopts a progressive learning strategy to gradually learn the structure of the OD matrix from a coarse to fine scale. OD-PGGAN utilizes multi-scale generators and discriminators to perform generation and discrimination tasks at different spatial resolutions. OD-PGGAN introduces a geography-based upsampling and downsampling algorithm to maintain the geographical significance of the OD matrix during spatial resolution transformations. The results demonstrate that the proposed OD-PGGAN can generate a large-scale synthetic OD matrix with 1024 nodes that have the same distribution as the real sample and outperforms two classical methods. The OD-PGGAN can effectively provide reliable synthetic data for transportation applications.

Keywords: origin–destination matrix; generative adversarial networks; intelligent transportation systems; gravity model; radiation



Academic Editors: Wolfgang Kainz and Hartwig H. Hochmair

Received: 15 January 2025

Revised: 8 April 2025

Accepted: 10 April 2025

Published: 14 April 2025

Citation: Yuan, Z.; Chen, X.; Chen, B.; Luo, Y.; Zhang, Y.; Teng, W.; Zhang, C. Generating Large-Scale Origin–Destination Matrix via Progressive Growing Generative Adversarial Networks Model. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 172.
<https://doi.org/10.3390/ijgi14040172>

Copyright: © 2025 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Transportation plays a vital role in human society. Advancements in both transport systems and infrastructure have a growing influence on our society, economy, and environment. Over the past few decades, the volume and density of vehicles have increased significantly, leading to increased accidents and congestion. The expansion of traffic infrastructure and new roads can alleviate some of these problems but is insufficient to fully address the escalating demand for mobility. With technological advancements, particularly in computer science, intelligent transportation systems (ITS) and a series of new methods and applications have been proposed for traffic management. These innovations improve

traffic safety and efficiency and have garnered widespread attention from governmental organizations and scientific communities. In the application of ITS, understanding the dynamics of traffic demand is an important topic, as it helps us to optimize the movement of vehicles and improve traffic flow. The origin–destination (OD) matrix is a matrix that describe the total number of traffic flows between regions and serves as crucial input for transportation modeling [1,2].

Traditionally, to construct the OD matrix, traffic flow data were obtained through census data, local travel surveys, vehicle counts collected through loops or manual, deploying sensors on the road network, etc. With advances in information and communication technology (ICT) and location-based services (LBS), spatiotemporal big data have opened up new possibilities for collecting traffic flow data. Spatiotemporal big data include probe car data, cellular-based mobile phone signaling data, social media data, public transportation smart card data, etc. [3]. However, existing datasets still face limitations, particularly in terms of high costs and privacy issues. Collecting data in a new city often requires a significant investment of time and resources, such as the expense of deploying sensors and recruiting volunteers. Moreover, since traffic mobility data contain the approximate whereabouts of individuals, it raises increasing privacy considerations [4,5].

To overcome the limitations of high costs and privacy concerns associated with collecting real-world traffic mobility data, researchers have turned to synthetic data generation as a promising solution [6–8]. Synthetic data are generated by model and have the same statistical distribution as real data. This enables synthetic data to serve as a viable substitute for real data in both research and application by preserving statistical characteristics while safeguarding sensitive information [9]. Synthetic data help us to address privacy concerns and compensate for incomplete, scarce, or biased datasets [10,11]. In this context, the OD matrix generation task has garnered considerable critical attention. The OD matrix generation task aims to construct synthetic OD matrix that have the same distribution as the real OD matrix [12]. OD matrix generation is particularly important in two scenarios. First, in cities where traffic flow data are limited or absent due to the high cost of deploying sensors or conducting surveys, OD matrix generation allows researchers to obtain usable OD matrix in the absence of traffic flow data. Second, in cases where real mobility data (e.g., mobile phone signaling data, GPS floating car data) are available but not publicly accessible due to privacy regulations, the synthetic OD matrix can serve as a substitute in research while protecting individuals' privacy. Therefore, using OD matrix synthetic data helps us to address the challenge of obtaining OD matrix data in ITS research, especially in cases where real-world datasets are either difficult to obtain because of high costs or unsuitable for public use due to privacy concerns.

Previous studies on OD matrix generation generally fall into two primary categories: traditional physics-based methods and data-driven machine learning methods [13]. Physics-based methods apply physical laws to model traffic flow, such as the gravity model, radiation model, intervention opportunity model, etc. The second category is data-driven machine learning methods. With the development of machine learning, models such as random forest [14], Kalman filters [15], and neural networks [16,17] have been widely used in traffic flow simulation. These machine-learning-based models construct complex fitting functions to capture the complex nonlinear relationship between traffic flow and urban features such as transportation networks, urban land use, and points of interests (POIs), while also capturing the spatial relationships among different urban regions. As a result, machine-learning-based models often achieve better performance than physics-based models [18,19]. However, existing research still has limitations, particularly in the synthetic data generation task. Physics-based models have fewer parameters and often ignore multiple factors that influence movement behavior (e.g., transportation facilities

and urban built environment), resulting in suboptimal performance. On the other hand, machine-learning-based models primarily focus on fitting training data instead of capturing the intrinsic distribution of movements, which lead to a poor generalization capability [13]. Neither method is an effective solution for the OD matrix generation task.

In recent years, with the advancement of deep learning technologies, the emergence of Generative Adversarial Networks (GAN) has demonstrated significant potential for diverse applications in fields such as computer vision, speech and language processing, and ITS [20,21]. Compared to other deep learning algorithms, GAN have the advantage of generating synthetic data, which are indistinguishable from real data by learning the distribution of real data [22]. Therefore, GAN are widely used in synthetic data generation tasks across various fields. In the field of ITS, GAN are applied in tasks like autonomous driving data generation, individual trajectory generation, and OD matrix generation. In the OD matrix generation task, researchers have developed various improved GAN architectures to simulate and generate urban traffic flow. Studies have shown that GAN can effectively generate synthetic traffic flow data and achieve better performance compared to physics-based methods and machine-learning-based methods [13]. GAN based method effectively addresses issues such as data insufficiency and privacy concerns associated with using real data.

Although GAN-based methods can effectively generate the synthetic OD matrix, research on using GAN to generate large-scale OD matrix still remains limited. Due to rapid urbanization, urban areas have expanded significantly. A large-scale OD matrix with more nodes can capture urban traffic flows at a finer scale, which is important in the study of fine-scale urban studies and ITS research. Using GAN to generate large-scale OD matrix presents several challenges. There are two challenges: How do we generate OD matrix with thousands of nodes? In the OD matrix generation task, the output size of GAN directly determines the scale of the OD matrix. Current GAN-based studies can typically only generate OD matrix with tens of nodes [23]. A large-scale OD matrix generation task means handling the topological connections between millions of nodes, which significantly increases the complexity of the training process. Another challenge is how do we capture the implicit spatial relationships in OD matrix? OD matrix elements inherently encode spatial dependencies and geographical correlations. Directly applying image-based processing algorithms may fail to preserve these implicit geographical relationships.

To address the above challenges, this study proposes the Origin–Destination Matrix Progressive Growing Generative Adversarial Networks (OD-PGGAN) model for the OD matrix generation task to address the challenge of generating a largescale OD matrix. Progressive Growing Generative Adversarial Networks (PGGAN) have demonstrated superior performance in the task of generating high-resolution images. OD-PGGAN have adapted the PGGAN architecture for the large-scale OD matrix generation task. This study has designed a multi-scale generator and discriminator to generate and discriminate the OD matrix at different spatial resolutions. OD-PGGAN start the training from the low-resolution OD matrix and then progressively increase the spatial resolution of the OD matrix. The training strategy allows the networks to first discover the coarse-scale structure of the OD matrix and then shift attention to finer-scale details. This study designed a geography-based upsampling and downsampling algorithm to replace the upsampling and downsampling algorithm which is used in the field of image processing. By employing new upsampling and downsampling algorithms, this paper constructs mappings between the OD matrix of different spatial resolutions based on the hidden spatial relationships in the OD matrix among different regions. This approach helps ensure that the generated matrix maintains realistic geospatial semantics during progressive training. Based on mobile phone signaling data from Shanghai, this study has successfully trained an OD-

PGGAN model to generate a large-scale synthetic OD matrix with 1024 nodes and validated that OD-PGGAN models have superior performance compared to the baseline models.

The main contributions of this study are as follows. Firstly, the proposed OD-PGGAN model is capable of generating a large-scale OD matrix with thousands of nodes. Compared to traditional models, GAN-based models can produce more diverse synthetic OD matrix that have the same distribution as a real sample, demonstrating the potential of GAN-based models for large-scale OD matrix generation tasks. Secondly, this study introduces a geography-based upsampling and downsampling algorithm that can capture the inherent spatial relationships between the OD matrix at different spatial resolutions. This approach ensures that the progressive training process preserves the spatial relationship of the OD matrix, allowing for a more accurate representation of real-world traffic flow patterns across varying scales.

The structure of this paper is as follows: Section 2 reviews related work on GAN models and OD matrix generation methods. Section 3 introduces the methodology, covering key concepts, the definition of the research problem, and the architecture of OD-PGGAN and its improvements, including the multi-scale generators and discriminators and geography-based upsampling and downsampling algorithms. Section 4 presents the experimental setup, including experimental data, baseline models, and evaluation metrics. Section 6 presents the experiments results, demonstrating the superior performance of OD-PGGAN in generating large-scale OD matrix task. Section 6 summarizes the main content of the research and discusses future research directions.

2. Literature Review

This section provides a detailed review of the related works on GAN and OD matrix generation.

2.1. Generative Adversarial Networks

Goodfellow et al. first introduced the framework of GAN in 2014 [24]. In their proposed framework, they trained the model using only backpropagation and dropout algorithms and generated samples from the model using only forward propagation. The results showed that GAN have the potential to compete with the other generative models. GAN consist of two components: a generator and a discriminator. The generator attempts to learn the distribution of real samples and generate fake samples, while the discriminator, typically a binary classifier, is used to distinguish the fake samples from real samples as accurately as possible. The goal of generator and discriminator are opposite to each other. The goal of the generator is to generate fake samples that resemble real samples closely enough to fool the discriminator. Conversely, the discriminator aims to distinguish real and fake samples accurately. GAN are trained using a minimax optimization framework, seeking a Nash equilibrium between the two components [25,26]. At this equilibrium, the generator can effectively learn the data distribution and generate synthetic data which the discriminator cannot distinguish from real samples.

Since the introduction of GAN, researchers have developed numerous enhanced architectures such as Conditional GAN (CGAN) [27], Deep Convolutional GAN (DCGAN) [28], Semi-Supervised GAN (SGAN) [29], Wasserstein GAN (WGAN) [30], Graph Representation Learning with GAN (GraphGAN) [31], and Diffusion GAN [32]. These refined architectures effectively mitigate challenges in GAN training, notably mode collapse and non-convergence. GAN exhibit superior performance across applications including image synthesis, video and audio synthesis, natural language processing, medical and healthcare, security, and various domains [33].

With the advancement of GAN, several studies have begun to pursue higher resolution and clarity of images. Consequently, numerous novel GAN architectures have proposed for high-resolution image generation, conversion, and editing applications, such as Progressive Growing GAN (PGGAN) [34], StyleGAN [35], BigGAN [36], Self-Attention GAN (SAGAN) [37], StyleSwin [38], and so on. These advancements have significantly improved the fidelity and diversity of generated images, benefiting applications in image synthesis [39], medical imaging [40], geoscience [41], and remote sensing [42].

Beyond computer vision, GAN have been widely applied in synthetic data generation, particularly in scenarios where real data collection is constrained by privacy concerns or data scarcity. GAN can generate synthetic data that preserve the statistical characteristics of the original data while safeguarding sensitive information, ensuring that data can be securely utilized for research and analytical purposes [9,43]. Various GAN-based models have been developed to generate realistic yet privacy-preserving datasets in domains such as medical imaging, mobility patterns, and sensor data simulation. In ITS research, GAN have demonstrated significant application potential and have been leveraged for tasks including traffic flow modeling, anomaly detection, and autonomous driving perception [21]. In recent years, the use of synthetic data has gained widespread attention in ITS research. Recent studies have employed GAN-based methods for trajectory generation [44] and OD matrix generation to address challenges related to privacy protection and data scarcity. The related work of GAN based OD matrix generation studies will be further discussed in Section 2.2.

2.2. OD Matrix Generation

The OD matrix generation task refers to constructing OD matrix based on the known geographical information through models, without any OD flow information available. Specifically, by giving a set of geographic locations, along with the socio-geographic attributes of these regions (e.g., population and land use) and the interaction characteristics between different regions (e.g., distance and spatial adjacency), researchers are able to generate traffic flow (e.g., population and vehicle) between regions and construct the OD matrix. In addition to generation, this issue is also related to the traffic flow inference, estimation, imputation, construction, and prediction [12]. Existing research on OD matrix generation could be summarized into two categories.

The first category methods are physics-based methods that typically use physical laws to simulate the traffic flow, primarily including the gravity model, radiation model, and intervening opportunities model, among others. For instance, the gravity model often likens the movement of populations between two regions to the gravitational attraction between celestial bodies, while the radiation model equates population movement to the radiation and absorption processes in physics. Recent research often employs more complex physical models or considers additional variables to simulate traffic flow [45–48]. The advantage of physical models lies in their ability to generate traffic flow OD matrix based on preset model mechanisms, minimal data, and specific parameters. However, physical models struggle to capture the complex relationships between urban spatial structures and traffic flows, often resulting in significant deviations from reality and poor performance. The second category is machine-learning based models. With the advancement of machine learning, contemporary models such as random forests, neural networks, and graph neural networks can model the complex nonlinear relationships between variables. Compared to physical models, researchers incorporate more geographical features such as land use types, POI density, weather conditions, and the interactions between geographical regions, significantly improving the accuracy of model outputs [49–51].

Recently, using generative algorithms for data production has become a major research focus. Scholars have also conducted a series of studies on generating traffic OD matrix using generative algorithms. Mauro et al. proposed the Mobility Generative Adversarial Network (MoGAN), which is based on the architecture of DCGAN [23]. The results showed that the realism of the OD matrix generated by MoGAN outperforms that of physical models, demonstrating the potential of GAN in the OD matrix generation task. Rong et al. construct an ODGAN model to generate the OD matrix which integrates physics laws and machine learning methods [13]. Li et al. proposed a Conditional Wasserstein Generative Adversarial Network with Gradient Penalty (CWGAN-GP) to predict the passenger demand OD matrix [52]. Rong et al. proposed a cascaded graph denoising diffusion OD matrix generation method (DiffODGen) which comprised a two step diffusion model for generating the OD matrix in new urban [53].

These methods still have limitations in the OD matrix generation task. The gravity and radiation model are designed to generate a single matrix, while machine learning models tend to directly fit the training data, which lead to their low generalization capability. Using GAN models helps us to address these issues, but there has not yet been an efficient method on generating a large-scale OD matrix with thousands of nodes, and existing studies can only generate the OD matrix with only tens of nodes.

This study proposed OD-PGGAN to explore the potential of using GAN to generate a large-scale OD matrix. This paper introduced a progressive growing network structure to generate OD networks with 1024 nodes. This paper employed multi-scale generators and discriminators to generate the synthetic OD matrix at different spatial resolutions. The geography-based upsampling and downsampling algorithm is designed to capture the geographical topology between nodes in the OD matrix at different spatial resolutions.

3. Methodology

3.1. Preliminary

This section provides a systematic introduction to the necessary definitions and necessary notations and presents the problem formulation for large-scale OD matrix generation.

3.1.1. Definitions

Definition 1 (Study Region). *The study region refers to the area where the OD matrix generation task is conducted, denoted as R .*

Definition 2 (Region). *A region r describes a smaller area contained within the study region, denoted as $r \in R$. The study region R , according to a specific partitioning algorithm, is divided into n non-overlapping regions, denoted as follows,*

$$R^{(n)} = \bigcup_{i=1}^n r_i | \forall i, j \in \{1, 2, \dots, n\}, i \neq j \Rightarrow r_i \cap r_j = \emptyset \quad (1)$$

Definition 3 (Trajectory). *A trajectory is defined as the path that an individual follows during a single movement in geographical space, represented by a series of chronologically ordered points, where each point contains geographical coordinates and a timestamp. For example, a trajectory T can be represented as $T = \{(x_1, y_1, t_1), (x_2, y_2, t_2), \dots, (x_i, y_i, t_i)\}$. In this study, the point (x_i, y_i) can be represented by the region $r_{(x_i, y_i)}$ in which the point is located. Similarly, trajectory can also be represented as $T = \{(r_{(x_1, y_1)}, t_1), (r_{(x_2, y_2)}, t_2), \dots, (r_{(x_i, y_i)}, t_i)\}$. Specifically, the trajectory is denoted as $T = \{(r_{org}, t_1), (r_{dst}, t_i)\}$, it indicates that this trajectory departs from the origin region r_{org} and terminates at the destination region r_{dst} .*

Definition 4 (OD Flow). The directed OD flow $F_{r_{org}, r_{dst}}$ describes the cumulative mobility flow of all trajectories T that depart from the origin region r_{org} and terminates at the destination region r_{dst} over a specified period.

Definition 5 (OD Matrix). The OD matrix M is an $n \times n$ matrix that includes the OD flows between every pair of regions in $R^{(n)}$, defined as follows in Equation (2). Each node $m_{i,j}$ of the matrix M represents the OD flow F_{r_i, r_j} that departs from origin region r_i to destination region r_j .

$$M = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,n} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n,1} & m_{n,2} & \cdots & m_{n,n} \end{bmatrix} \quad (2)$$

Definition 6 (OD Network). The OD matrix M can be represented as a weighted directed graph $G = (V, E, \omega)$ from the network perspective. V is the set of nodes, where each node represents a region, E is the set of directed edges between every two regions, and ω_{r_i, r_j} is the weight function on the edges, denoting the OD flow F_{r_i, r_j} .

Definition 7 (Spatial resolution). Spatial resolution refers to the level of detail captured in the OD matrix, similar to the concept used in remote sensing. For a given study region, higher spatial resolution means that the study region is divided into a subset of smaller regions, with each region representing a fine spatial unit, allowing for a more detailed representation of traffic flow. The spatial resolution can be represented as $R^{(n)}$, where n denotes the numbers of regions. Specifically, an OD matrix with n larger than 1000 is considered a large-scale OD matrix.

3.1.2. Problem Formulation

Definition 8 (OD Matrix Generation Task). Given the study region $R^{(n)}$ and a series of historical OD matrix data $\mathcal{X}_{real} = \{M_1, M_2, \dots, M_n\}$, the OD matrix generation task involves developing a model that can produce a set of synthetic OD mobility matrix $\mathcal{X}_{synthetic} = \{\hat{M}_1, \hat{M}_2, \dots, \hat{M}_m\}$. Specifically, when n is larger than 1000, the generation task is a large-scale OD matrix generation task.

3.2. Methodology

This section presents the framework of our proposed methodology, OD-PGGAN. To address the issue of large-scale OD matrix generation, we designed the OD-PGGAN, a deep learning architecture based on the PGGAN framework. OD-PGGAN employ multi-scale generators that progressively increase the spatial resolution of generated OD matrix from low to high, combined with multi-scale discriminators that correspond to different spatial resolutions. Additionally, geography-based upsampling and downsampling algorithms are incorporated to enhance the performance of the PGGAN model. Figure 1 illustrates the workflow of this study. Mobile phone signaling data are used as the data source. After performing stay point detection and trajectory aggregation, this study constructs the OD matrix using the sampling approach. The OD matrix is then divided into training and test datasets. The training dataset is used to train OD-PGGAN, while the test dataset is utilized for validation. Additionally, the gravity and radiation models are fitted for comparison.

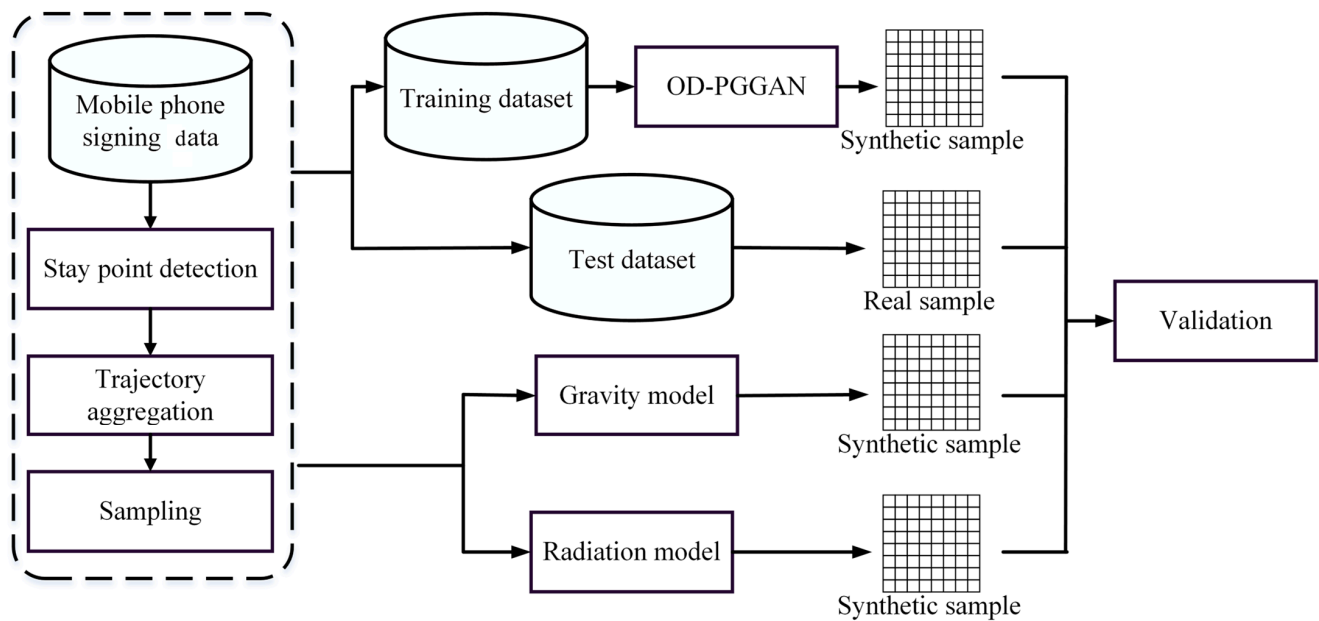


Figure 1. Research workflow of OD-PGGAN for large-scale OD matrix generation task.

3.2.1. Network Structure

This section provides a detailed introduction to the framework of our proposed OD-PGGAN. As visualized in Figure 2, OD-PGGAN leverage the PGGAN architecture and are designed for the task of high-spatial-resolution OD matrix generation. The training strategy of OD-PGGAN is similar to the PGGAN, where training begins with the low-spatial-resolution OD matrix and progressively adds new structures to the network to gradually increase the spatial resolution.

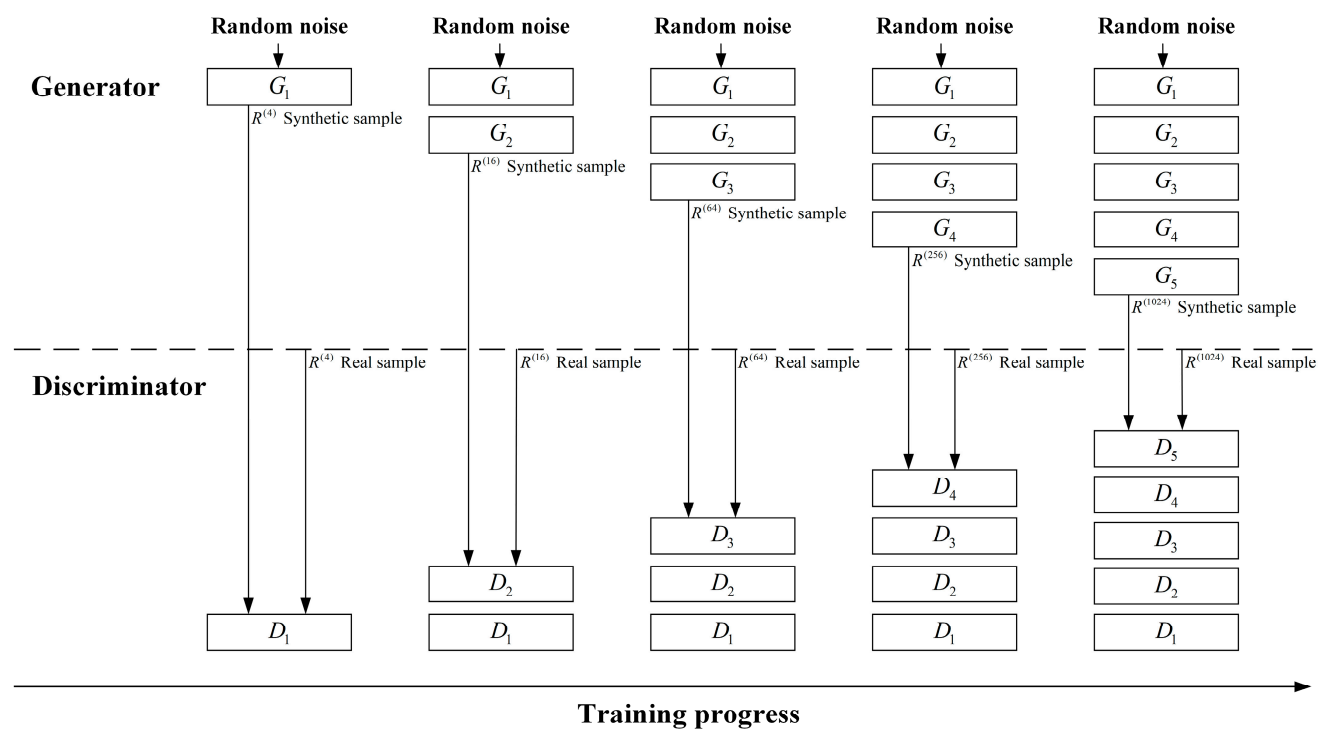


Figure 2. Architecture of OD-PGGAN.

This study designed multiple multi-scale generators G_1 , G_2 , G_3 , G_4 , and G_5 . Corresponding to multi-scale generators, we designed multiple multi-scale discriminators D_1 ,

D_2 , D_3 , D_4 , and D_5 , to generate and discriminate the OD matrix at spatial resolutions of $R^{(4)}$, $R^{(16)}$, $R^{(64)}$, $R^{(256)}$, and $R^{(1024)}$, respectively. The generators and discriminators are in a mirror structure, growing synchronously as training progresses. The design of the multi-scale generators and discriminators is based on the principles of PGGAN, where the complex mapping from low resolution to high resolution is easier to gradually learn step by step. The crucial difference is that OD-PGGAN introduce a geography-based upsampling and downsampling algorithm in the training progress, ensuring that the geographical significance of the OD matrix is preserved during the spatial resolution transformations. This progressive training strategy allows us to capture the coarse structure of the OD matrix and gradually refine the details as the spatial resolution increases.

The goal of the generator is to map the random input noise z to an OD matrix of a specific spatial resolution. In the initial stage of training, G_1 maps the random noise z to a 4×4 matrix at the spatial resolution $R^{(4)}$. As training stabilizes, new layer structures are added to the network. The low-spatial-resolution synthetic OD matrix is upsampled via a designed geography-based upsampling algorithm and fed to the next level of the generator to produce an OD matrix with a higher spatial resolution, ultimately reaching the target resolution $R^{(1024)}$, which generates a synthetic OD matrix with 1024 nodes.

To correspond with the generator's output spatial resolutions, OD-PGGAN designed multi-scale discriminators that grow in synchrony. Each scale's discriminator evaluates the OD matrix at a corresponding spatial resolution. Specifically, OD-PGGAN employ a geography-based downsampling algorithm on the real OD matrix to create the OD matrix pyramid. This OD matrix pyramid serves as the training data for discriminators, which enables each discriminator to distinguish the real OD matrix from the synthetic OD matrix across different spatial resolutions.

3.2.2. Multi-Scale Generators and Discriminators

This study designed five generators, G_1 , G_2 , G_3 , G_4 , and G_5 to generate the synthetic OD matrix at different spatial resolutions, and each generator outputs an OD matrix with a specific spatial resolution at $R^{(4)}$, $R^{(16)}$, $R^{(64)}$, $R^{(256)}$, and $R^{(1024)}$, respectively. Therefore, our generator can be represented as a tuple $G = \{G_1, G_2, \dots, G_n\}$; as additional network components are added to the generator, the output resolution is determined by the final generator in the sequence. For example, if the generator tuple is defined as $G = \{G_1, G_2\}$, the output size of the matrix is 16×16 . The use of multi-scale generators helps us to gradually learn the structure of the OD matrix. Corresponding to the generator, five discriminators, D_1 , D_2 , D_3 , D_4 , and D_5 were designed, with each discriminator designed to distinguish the real sample and the synthetic sample at spatial resolutions of $R^{(4)}$, $R^{(16)}$, $R^{(64)}$, $R^{(256)}$, and $R^{(1024)}$, respectively. Similarly to generators, discriminator can be represented as a tuple $D = \{D_1, D_2, \dots, D_n\}$. For example, if the discriminator tuple is defined as $D = \{D_1, D_2\}$, then discriminator D will accept the input size of a matrix of 16×16 and perform the discrimination task across two different spatial resolutions.

As shown in Figure 3, in the initial stage, the training process begins with the low-spatial-resolution generator $G = \{G_1\}$ and discriminator $D = \{D_1\}$ at the spatial resolution $R^{(4)}$. G_1 takes random noise z as the input after a series of transposed convolution operations to generate the synthetic OD matrix at $R^{(4)}$. Simultaneously, based on the real sample, this study downsamples the real sample from $R^{(1024)}$ to $R^{(4)}$ using a geography-based downsampling algorithm. The synthetic sample and the downsampled real sample are then input into the corresponding discriminator D_1 . After the initial training phase, OD-PGGAN introduce new network structures in GAN to produce higher-spatial-resolution OD matrix. We define the generator $G = \{G_1, G_2\}$ and discriminator $D = \{D_1, D_2\}$ at the spatial resolution of $R^{(16)}$. In the generator G , G starts by accepting z as input. After

processing through G_1 , the output is a 4×4 matrix. This matrix is then passed through a geography-based upsampling process to a 16×16 matrix and used as input for G_2 . Subsequently, after a series of transposed convolution operations, G_2 produces an output matrix at spatial resolution $R^{(16)}$. Correspondingly, based on the real sample, OD-PGGAN apply a geography-based downsampling algorithm to reduce the real sample from a resolution of $R^{(1024)}$ to $R^{(4)}$. The downsampled real data will be used for training the discriminator D_2 . In D_2 , both the real sample and synthetic sample are input to perform binary classification tasks at this specific resolution. Additionally, the synthetic sample undergoes geography-based downsampling to the $R^{(4)}$, and then passed through D_1 to determine whether the synthetic matrix still contains the coarse structure at a lower spatial resolution. As the training process progresses, new network structures are synchronously added to both G and D , gradually increasing the resolution of the OD matrix generated by OD-PGGAN until reaching the spatial resolution of $R^{(1024)}$.

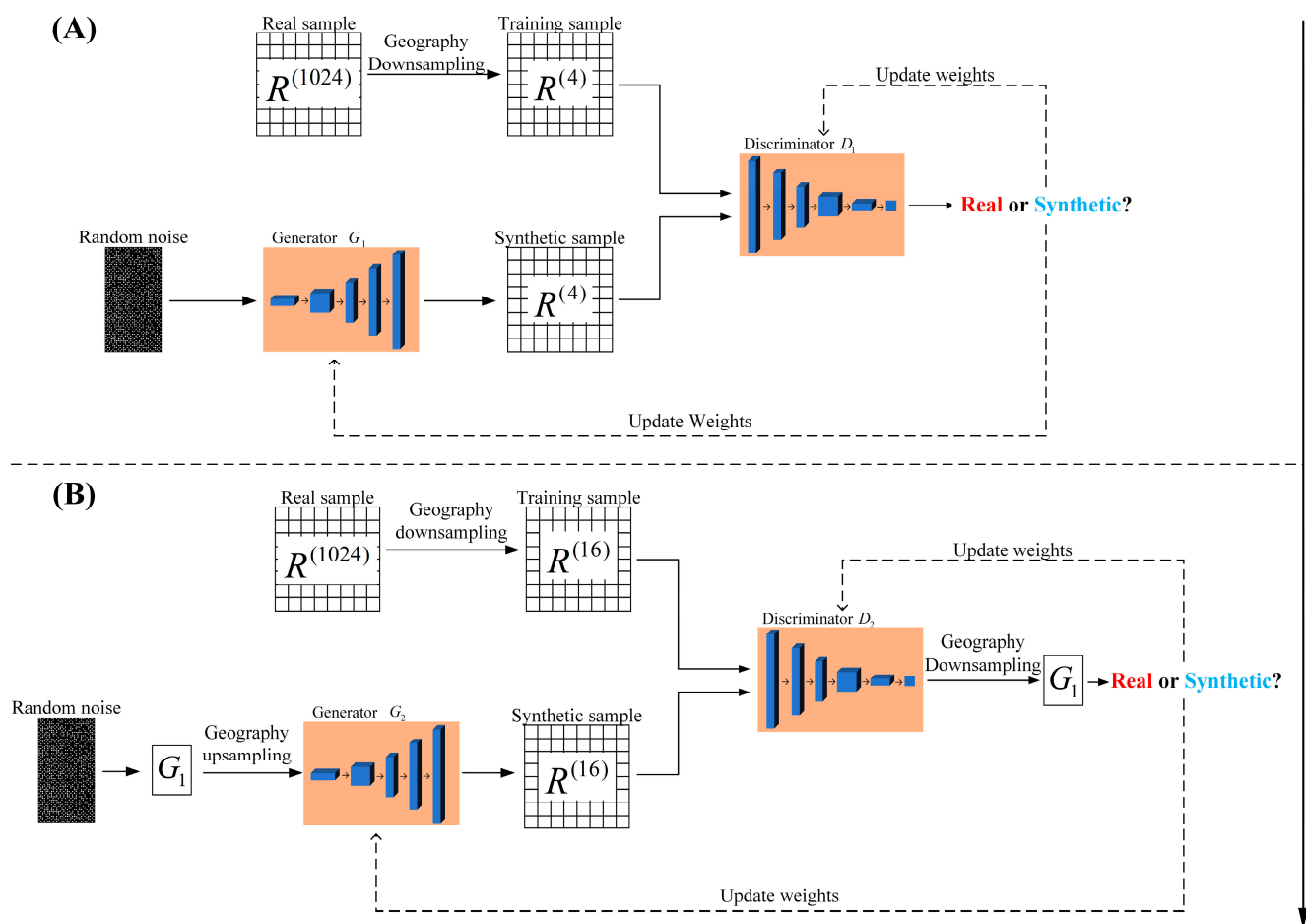


Figure 3. Example structures of multi-scale generators and discriminators. (A) The structure of the generator $G = \{G_1\}$ and discriminator $D = \{D_1\}$ at spatial resolution $R^{(4)}$. (B) The structure of the generator $G = \{G_1, G_2\}$ and discriminator $D = \{D_1, D_2\}$ at spatial resolution $R^{(16)}$.

3.2.3. Geography-Based Upsampling and Downsampling Algorithm

In the field of computer vision, upsampling and downsampling techniques are fundamental operations for image preprocessing [54]. These techniques adjust the resolution and size of an image by increasing or decreasing the number of pixels, with the aim of retaining the image's visual features based on specific sampling algorithms. However, when applied to the task of upsampling and downsampling an OD matrix, directly using image processing sampling algorithms presents limitations. Upsampling and downsampling

operations in an image matrix rely on the spatial relationships between pixels. In an image matrix, neighboring elements correspond to adjacent pixels in the actual image. Therefore, image processing sampling algorithms can leverage this explicit spatial relationship to preserve visual continuity and ensure smooth transitions between neighboring pixels during resolution adjustments.

In contrast, the OD matrix differs from the image matrix. Each element in the OD matrix represents flow information between specific geographic regions. For instance, neighboring elements in an OD matrix do not necessarily represent geographically adjacent areas. Consequently, directly applying image-based upsampling and downsampling algorithms to an OD matrix may lead to a loss of geographic context. Specifically, these algorithms may introduce incorrect spatial adjacency spatial relationship in the OD matrix, resulting in inaccurate allocation and adjustment of flow information between regions and leading to outcomes that do not reflect the actual geographic layout.

Therefore, for the large-scale OD matrix generation task, this study designed a geography-based upsampling and downsampling algorithm specifically designed for OD matrix spatial resolution adjustments. This algorithm performs sampling based on real spatial adjacency relationships, thereby preserving the geographic adjacency in the OD matrix across different scales. This approach ensures that the OD matrix retains its geographical significance during the upsampling and downsampling processes.

Given the study area with a low spatial resolution, denoted by $R^{(low)}$ and its corresponding OD matrix $M^{(low)}$, as well as a high spatial resolution $R^{(high)}$ with OD matrix $M^{(high)}$. For each low spatial resolution $r_i^{(low)}$, it can be represented by multiple high-spatial-resolution regions $r_{p_k}^{(high)}$, as shown in Equation (3):

$$\forall r_i^{(low)} \in R^{(low)}, \exists \{r_{p_1}^{(high)}, r_{p_2}^{(high)}, \dots, r_{p_k}^{(high)}\} \in R^{(high)} : r_i = \bigcup_{l=1}^k r_{p_l} \quad (3)$$

In the downsampling process, the OD matrix $M^{(low)}$ aggregates traffic flow based on geographic adjacency. For each element $M_{i,j}^{(low)}$ in $M^{(low)}$, the traffic flow is computed as Equation (4):

$$m_{i,j}^{(low)} = \sum_{r_p \in r_i} \sum_{r_q \in r_j} m_{p,q}^{(high)} \quad (4)$$

where $M_{i,j}^{(low)}$ represents the aggregated OD flow from region i to region j in the low-spatial-resolution matrix, while $M_{p,q}^{(high)}$ denotes the OD flow from region p to region q in the high-spatial-resolution matrix. $r_p \in r_i$ and $r_q \in r_j$ indicate that r_i and r_j are the sets of all high-spatial-resolution regions that belong to the downsampling regions i and j .

In the upsampling process, to transform a low-spatial-resolution OD matrix $M^{(low)}$ into a higher-spatial-resolution matrix $M^{(high)}$, each element $M_{p,q}^{(high)}$ is calculated as follows:

$$m_{p,q}^{(high)} = m_{i,j}^{(low)} \times \frac{\hat{m}_{p,q}^{(high)}}{\sum_{r_p \in r_i} \sum_{r_q \in r_j} \hat{m}_{p,q}^{(high)}} \quad (5)$$

where $\hat{M}_{p,q}$ represents the real traffic flow information between regions p and q calculated by a real sample. The term $\hat{M}_{p,q}^{(high)} / \sum_{R_p \in R_i} \sum_{R_q \in R_j} \hat{M}_{p,q}^{(high)}$ represents the proportion of flow in the high-spatial-resolution region relative to the total flow in the corresponding low-spatial-resolution region in the real sample.

Figure 4 illustrates the upsampling and downsampling processes between matrix with spatial resolutions of $R^{(4)}$ and $R^{(16)}$. For example, the low-spatial-resolution region $r_1^{(low)}$

maps to several high-spatial-resolution regions $r_1^{(high)}$, $r_2^{(high)}$, $r_5^{(high)}$, and $r_6^{(high)}$. Based on the mapping relationships between spatial regions, the traffic flow of $m_{1,1}^{(low)}$ is distributed to the elements $m_{1,1}^{(high)}$, $m_{1,2}^{(high)}$, $m_{1,5}^{(high)}$, $m_{1,6}^{(high)}$, $m_{2,1}^{(high)}$, $m_{2,2}^{(high)}$, $m_{2,5}^{(high)}$, $m_{2,6}^{(high)}$, $m_{5,1}^{(high)}$, $m_{5,2}^{(high)}$, $m_{5,5}^{(high)}$, $m_{5,6}^{(high)}$, $m_{6,1}^{(high)}$, $m_{6,2}^{(high)}$, $m_{6,5}^{(high)}$, and $m_{6,6}^{(high)}$. Conversely, during the upsampling process, the above elements are aggregated into $m_{1,1}^{(low)}$.

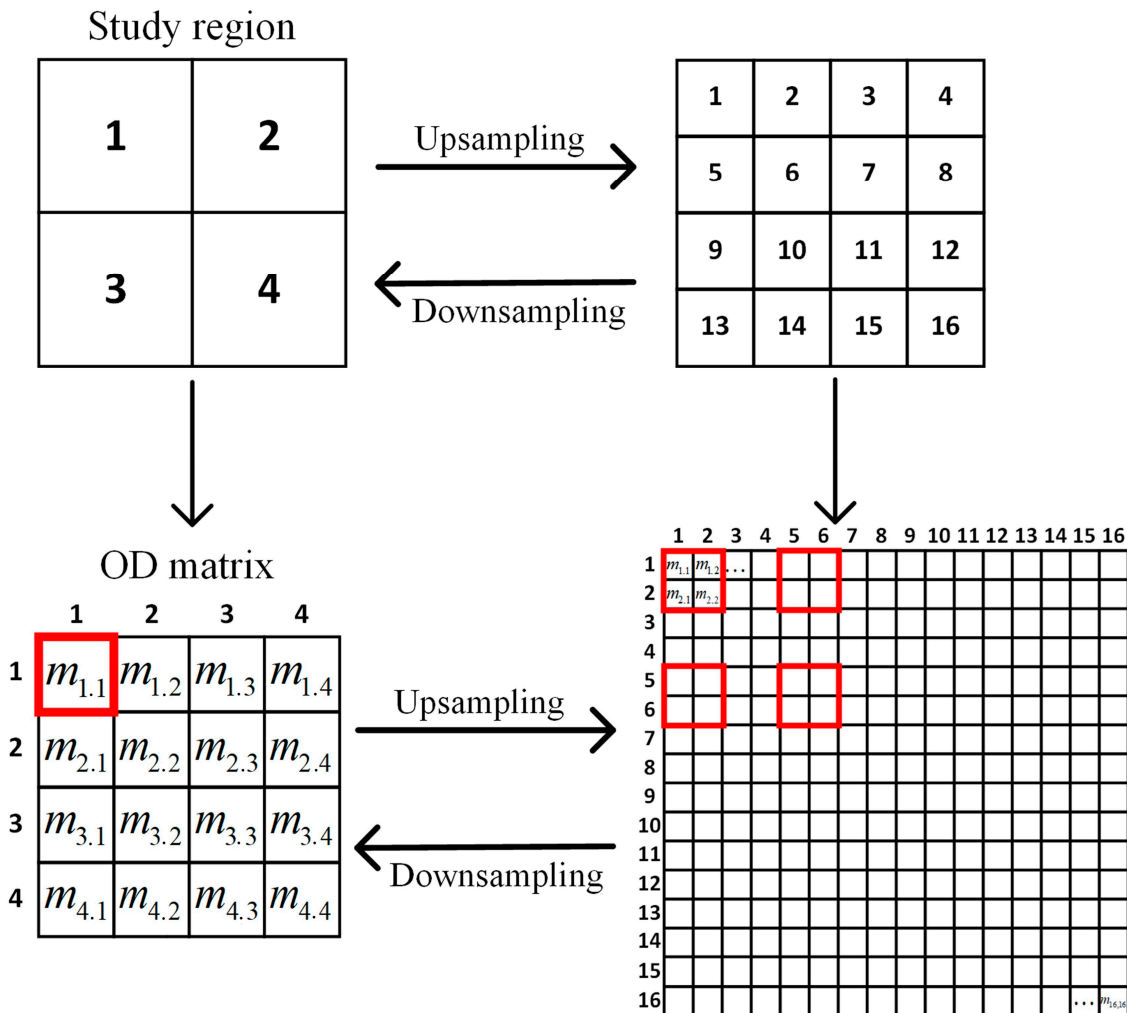


Figure 4. Example of geography-based upsampling and downsampling.

4. Experimental Setup

This study presents experiments on a real dataset of Shanghai mobile phone signaling data to answer the following research questions and assess whether the OD-PGGAN model can generate a large-scale synthetic OD matrix.

RQ1: Can our proposed OD-PGGAN model effectively generate the large-scale synthetic OD matrix?

RQ2: Compared to existing models, does the OD-PGGAN model demonstrate significant advantages?

4.1. Datasets

This study used mobile phone signaling data from Shanghai and its surrounding areas as the real-world dataset to evaluate the efficacy of the proposed OD-PGGAN model. This dataset was collected from a major telecommunication operator in China, covering data from 1 October to 31 October 2012, involving nearly 17 million users. These data record the

cell tower location information of users during activities such as phone calls, text messaging, switching between cell towers, and regular communication with cell towers. In Shanghai, the average coverage area of each tower is approximately 0.447 km^2 . To adhere to privacy standards, all user data were anonymized, ensuring there is no personally identifiable information. The anonymization process included the removal of direct identifiers such as telephone numbers, thereby safeguarding user privacy.

This study applied a stay point extraction algorithm to the mobile phone signaling data to identify individuals' meaningful stay points and moving behaviors and extract each individual's trajectory. Shanghai was divided into 1024 equally sized rectangular regions. Based on this regional division, this study further aggregated individual trajectories into OD flows. To ensure sufficient training data, this study adopted a non-replacement random sampling approach to construct the OD matrix. Specifically, for each day, 500,000 users were randomly sampled without replacement to construct an OD matrix, and this process was repeated iteratively until the remaining number of users was insufficient to form another complete OD matrix. In total, this sampling approach constructs 1020 OD matrix as the real sample over the 31-day period. Each OD matrix represents the OD flow information derived from 500,000 users sampled on a given day.

4.2. Validation

To address RQ1, we designed the validation approach based on Mauro [23]. The goal of the validation approach is to evaluate the realism of the synthetic OD matrix, and we need to verify whether the synthetic data generated by the model is capable of reproducing the variability of the OD matrix in the real sample dataset. If the distribution of differences among the OD matrix in the synthetic dataset is similar to the real sample dataset, it indicates that OD-PGGAN can effectively approximate the variability of the real sample dataset.

We split 1020 real sample matrix into a training dataset (820 matrix) and a test dataset (200 matrix). We trained OD-PGGAN on the training dataset and generated 200 synthetic OD matrix as the synthetic dataset. We then evaluated the realism of the model by computing the difference between each OD matrix in the synthetic dataset and each OD matrix in the test dataset. Additionally, we created a mixed dataset of 200 matrix, where half of the matrix were randomly chosen from the test dataset, while the other was randomly chosen from the synthetic dataset. We then computed the difference between every pair of OD matrix in the mixed dataset. The results from the mixed dataset provide further insights into whether the synthetic data exhibit similar variability to the real data.

To address RQ2, we compared our developed OD-PGGAN with two selected baseline models. Given the current lack of GAN-based models specifically designed for the high-resolution OD matrix generation task, we chose to compare our method at the spatial resolution of $R^{(1024)}$ with classical physics-based models: the gravity model and the radiation model.

4.3. Baseline Models

4.3.1. Gravity Model

The gravity model is inspired by Newton's law of gravitation. Zipf first introduced the gravity model to explain population mobility [55]. The gravity model is widely used in human mobility research to model mobility flow. In this research, we used the single-

constrained gravity model. The expected traffic flow F_{r_i, r_j} between the origin region r_i to the destination region r_j is calculated by the following equation:

$$F_{r_i, r_j} = O_i p_{r_i r_j} = O_i \frac{m_j^\beta f(r_{ij})}{\sum_k m_k^\beta f(r_{ik})} \quad (6)$$

where O_i is the population of location r_i , and $p_{r_i r_j}$ is the probability of observing a trip from r_i to r_j . Here, m_j is the population of the region r_j , and β is a parameter that controls the influence of the population on traffic flow. The deterrence function $f(r_{ij})$, represented as $f(r_{ij}) = d_{ij}^{-\alpha}$, describes the effect of distance on traffic flow, where d_{ij} is the distance between two region r_i and r_j , and α is a parameter controlling the impact of distance on flow. The term $\sum_k m_k^\beta f(r_{ik})$ represents the sum of the attractiveness of all potential destination regions of region r_i .

4.3.2. Radiation Model

The radiation model is a parameter free model introduced by Simini [56]. It generates the mobility flow between based on regional characteristics (e.g., population) and the intervening opportunities. The equation of the radiation model is shown below:

$$F_{r_i, r_j} = O_i \frac{1}{1 - m_i/M} \frac{m_i m_j}{(m_i + s_{ij})(m_j + s_{ij})} \quad (7)$$

where O_i is the total number of people leaving from region r_i . Here, m_i and m_j indicate the number of opportunities in region r_i and r_j . M is the sum of all the opportunities, and s_{ij} is the total number of opportunities in the circle of radius r_{ij} centered at region r_i .

4.4. Metrics

This study computes the difference between two OD matrix in two ways, by calculating an error metric and by computing the difference in edge weights. Specifically, we use three metrics to assess the difference: Normalized Root Mean Square Error (NRMSE), Common Part of Commuters (CPC), and Jensen–Shannon divergence (JSD).

The NRMSE is a min-max normalization of the Root Mean Square Error (RMSE), defined as follows:

$$NRMSE(A, B) = \frac{RMSE(A, B)}{\max(A, B) - \min(A, B)} \quad (8)$$

The RMSE is defined as follows:

$$RMSE(A, B) = \sqrt{\frac{1}{n^2} \sum_{i,j=1}^n (a_{ij} - b_{ij})^2} \quad (9)$$

where a_{ij} and b_{ij} are the elements at position (i, j) in the two OD matrix A and B , respectively, and n is the total number of elements in the matrix. The terms $\max(A, B)$ and $\min(A, B)$ represent the maximum and minimum values among all elements in matrix A and B .

The Common Part of Commuters (CPC), also known as the Sørensen–Dice index [57], is used to quantitatively measure the similarity between two OD matrix by calculating the overlap between the two matrix. CPC is a widely used metric in human mobility studies [58], with values ranging from 0 and 1. A CPC of 1 indicates a perfect match

between the synthetic OD matrix and the real OD matrix, while 0 means no agreement between the two matrix. Given two matrix, A and B , the CPC is calculated as follows:

$$CPC(A, B) = \frac{2 \sum_{i,j=1}^n \min(a_{ij}, b_{ij})}{\sum_{i,j=1}^n a_{ij} + \sum_{i,j=1}^n b_{ij}} \quad (10)$$

where a_{ij} and b_{ij} are the elements at position (i, j) in the matrix, respectively, and n represents the total number of elements in the matrix.

This study introduced Jensen–Shannon divergence (JSD) to measure the similarity between two matrix. When calculating the JSD between two matrix, the goal is to compare the distributions of edge weights or weight-distance metrics from a network perspective. JSD is a symmetric measure that quantifies the difference between two probability distributions. The JSD is defined as follows:

$$JSD(P \parallel Q) = \frac{1}{2}KLD(P \parallel M) + \frac{1}{2}KLD(Q \parallel M) \quad (11)$$

where P and Q are the probability distributions derived from matrix A and B , respectively, and $M = (P + Q)/2$ represents the average distributions of P and Q . The Kullback–Leibler divergence (KLD) is used to measure the difference between two distributions, and it is defined as follows:

$$KL(P \parallel Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \quad (12)$$

To obtain the probability distribution for the matrix, we define the weight adjacent matrix $\hat{A}$ as $\hat{A} = A / (d + \epsilon)$, where d is the distance matrix corresponding to the geographic distances between all pairs of nodes, and ϵ is a small residual term that prevents division by zero for diagonal elements of the matrix. And the distribution of matrix $\hat{A}$ is calculated by $P_{ij} = \hat{A}_{ij} / \sum_{ij} \hat{A}_{ij}$.

5. Result

Figure 5 shows the distribution of the CPC , $NRMSE$, and JSD metrics for the three datasets: the test set, mixed set, and synthetic set generated by OD-PGGAN. Figure 5 evaluates the internal variability of each dataset, reflecting the ability of the synthetic data generated by OD-PGGAN to replicate the statistical distribution of the real dataset. Across all five metrics, the distribution of the synthetic dataset demonstrates the same distribution as the test dataset. The mixed dataset, created by combining samples from both the test and synthetic datasets, exhibits a distribution that resembles both datasets. Table 1 presents the statistical characteristics of the test, mixed, and synthetic datasets. The results indicate a high degree of consistency among the three datasets. These results demonstrate that the generated synthetic data can effectively reproduce the distribution of the real OD matrix dataset, preserving its internal variability.

Considering that physics-based models generate only a single OD matrix. To ensure a fair comparison with OD-PGGAN, which generate multiple OD matrix, we adopt these models to generate multiple OD matrix as well. Specifically, we used all users' data for each date to construct a daily OD matrix and then fit the gravity and radiation models based on the daily OD matrix, which allowed us to obtain 31 OD matrix for each model. Figure 6 illustrates the distribution of CPC , $NRMSE$, and JSD metrics for the four datasets: the test set, synthetic set generated by OD-PGGAN, synthetic set generated by the gravity model, and synthetic set generated by the radiation model. Across all five metrics, the distribution

of the OD-PGGAN-generated dataset closely aligns with the test dataset, suggesting that the synthetic OD matrix effectively preserve the statistical characteristics and variability of real OD matrix. In contrast, the gravity and radiation models exhibit distinct distributions, indicating that physics-based methods struggle to capture the heterogeneity of the real dataset.

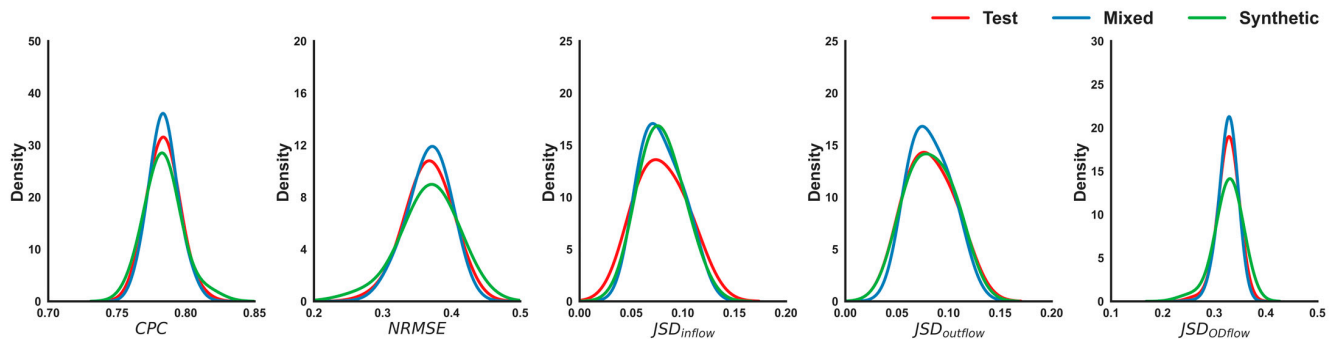


Figure 5. Distributions of *CPC*, *NRMSE*, and *JSD* for pairwise comparisons in the test dataset, mixed dataset, and synthetic dataset generated by OD-PGGAN.

Table 1. Statistics of *CPC*, *NRMSE*, and *JSD* in the test set, mixed set, and synthetic set.

	<i>CPC</i>	<i>NRMSE</i>	<i>JSD_{inflow}</i>	<i>JSD_{outflow}</i>	<i>JSD_{ODflow}</i>
Test Sample	0.784	0.367	0.079	0.082	0.327
Mixed Sample	0.783	0.364	0.077	0.079	0.322
Synthetic Sample	0.782	0.366	0.076	0.081	0.320

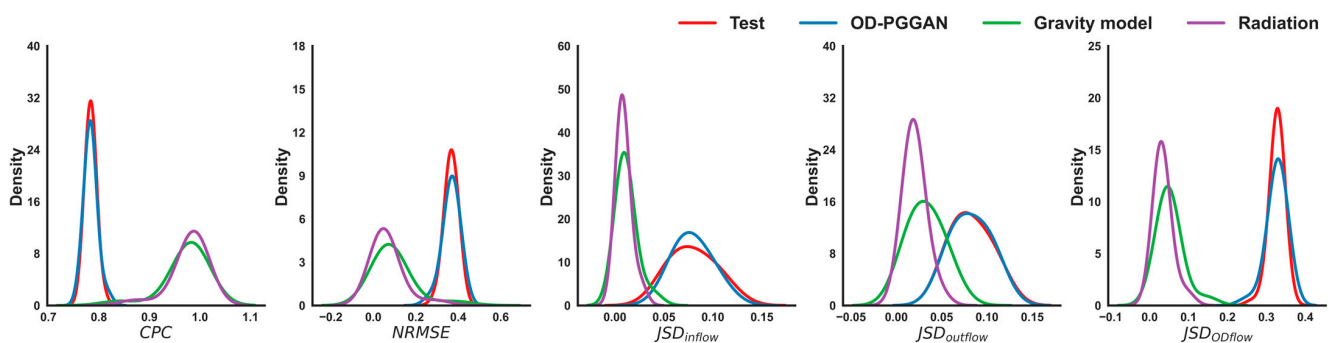


Figure 6. Distributions of *CPC*, *NRMSE*, and *JSD* for pairwise comparisons in the test set, synthetic set generated by OD-PGGAN, the gravity model, and the radiation model.

Table 2 presents the statistical characteristics, and the results indicate that OD-PGGAN achieve *CPC* and *NRMSE* values that are nearly identical to those of the test dataset, confirming the model's ability to generate OD matrix that reflect real-world mobility patterns. In contrast, the gravity and radiation models tend to generate highly similar matrix, which is reflected in their higher *CPC* values and lower *NRMSE* and *JSD* values. This indicates that the OD matrix dataset generated by physics-based models is insufficient to capture the variability between matrix in the real dataset. On the other hand, the synthetic dataset generated by OD-PGGAN effectively captures this variability, demonstrating that GAN-generated OD matrix better align with the real-world distribution.

Table 2. Statistics of *CPC*, *NRMSE*, and *JSD* for pairwise comparisons in the test set, OD-PGGAN, the gravity model, and the radiation model.

	<i>CPC</i>	<i>NRMSE</i>	<i>JSD_{inflow}</i>	<i>JSD_{outflow}</i>	<i>JSD_{ODflow}</i>
Test Sample	0.784	0.367	0.079	0.082	0.327
OD-PGGAN (synthetic sample)	0.782	0.366	0.076	0.081	0.320
Gravity Model	0.973	0.093	0.012	0.032	0.052
Radiation Model	0.979	0.066	0.009	0.020	0.035

Figure 7 shows the error distributions between the generated synthetic OD matrix and the real matrix for the proposed OD-PGGAN model, the gravity model, and the radiation model across the *CPC*, *NRMSE*, and *JSD* metrics. For these metrics, lower values of *NRMSE* and *JSD* and higher values of *CPC* indicate better performance. OD-PGGAN outperform baseline models across all evaluation metrics. Furthermore, the error distributions for OD-PGGAN are narrower and more concentrated, demonstrating its superior consistency and reliability in OD matrix generation tasks compared to the baseline models.

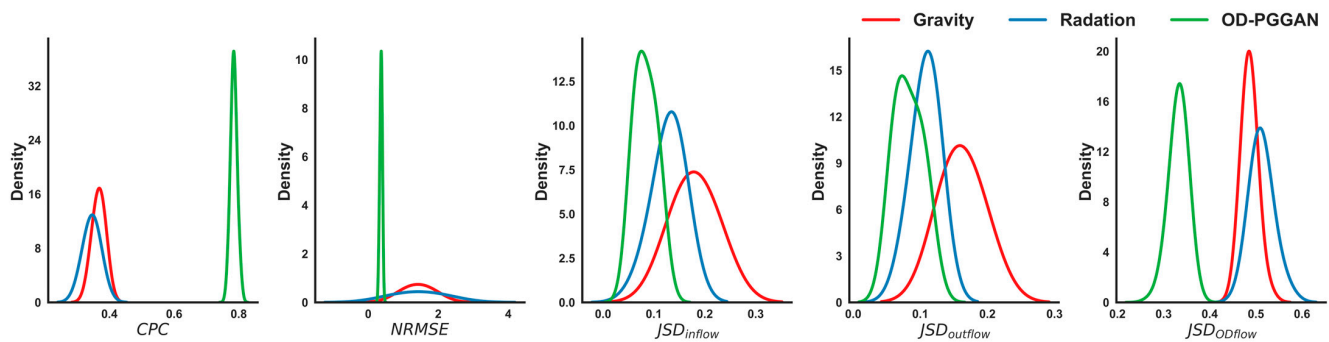
**Figure 7.** Distributions of *CPC*, *NRMSE*, and *JSD* for real and synthetic data across the gravity model, the radiation model, and OD-PGGAN.

Table 3 shows the errors between the synthetic OD matrix and the real OD matrix across several models. According to *CPC* and *NRMSE*, OD-PGGAN generate OD matrix that are more similar to the real sample, exhibiting smaller errors compared to the baseline models. Our results also show that the gravity model clearly outperforms the radiation model. To compute the improvement in the performance of OD-PGGAN with respect to the baseline models, for *JSD*, we define the quantity using Equation (13):

$$\Delta = \left(\frac{JSD^{(ODPGGAN)} - JSD^{(baseline)}}{JSD^{(baseline)}} \right) \times 100\% \quad (13)$$

According to the *JSD* metric, we found that OD-PGGAN significantly improve upon the baseline models. Specifically, the inflow *JSD* of OD-PGGAN improved by 56.1% compared to the gravity model and 38.8% compared to the radiation model. For outflow *JSD*, OD-PGGAN achieve an improvement up to 50.0% compared to the gravity model and 25.0% over the radiation model. In terms of OD flow *JSD*, OD-PGGAN demonstrate a 33.1% improvement over the gravity model and a 36.5% improvement on the radiation model.

Table 3. Statistics of *CPC*, *NRMSE*, and *JSD* across the gravity model, the radiation model, and OD-PGGAN.

	<i>CPC</i>	<i>NRMSE</i>	<i>JSD_{inflow}</i>	<i>JSD_{outflow}</i>	<i>JSD_{ODflow}</i>
Gravity Model	0.367	1.423	0.180	0.162	0.486
Radiation Model	0.346	1.435	0.129	0.108	0.512
OD-PGGAN	0.783	0.366	0.079	0.081	0.325

To ensure an adequate training dataset, this study adopts a sampling approach to construct OD matrix. To assess the bias introduced by the number of sampled users in OD matrix construction, we generate OD matrix using different sample sizes while applying the same sampling methodology. Table 4 presents the statistics of *CPC*, *NRMSE*, and *JSD* for OD matrix constructed using different sample sizes. It is important to note that when using the entire dataset to construct OD matrix, the results reflect variations across different days within the month rather than biases introduced by sampling. The results indicate that as the sample size increases, the *CPC* value also increases. This trend occurs because larger sample sizes result in fewer OD matrix, thereby reducing sampling bias. *NRMSE* also exhibits variations across different sample sizes. The inflow *JSD* remains relatively stable regardless of the sample size. The outflow *JSD* gradually stabilizes as the sample size increases, while the OD flow *JSD* decreases progressively with increasing sample size. These findings suggest that the OD matrix constructed using the sampling approach in this study can approximate the distribution characteristics of the full dataset to a certain extent.

Table 4. Statistics of *CPC*, *NRMSE*, and *JSD* for OD matrix constructed using different sample sizes.

	<i>CPC</i>	<i>NRMSE</i>	<i>JSD_{inflow}</i>	<i>JSD_{outflow}</i>	<i>JSD_{ODflow}</i>
100,000	0.772	0.368	0.083	0.086	0.342
200,000	0.775	0.375	0.081	0.085	0.342
500,000 (ours)	0.783	0.367	0.079	0.082	0.326
1,000,000	0.785	0.362	0.080	0.081	0.328
All users	0.792	0.357	0.081	0.082	0.311

6. Conclusions and Future Research

The OD matrix describes traffic flow information within urban regions and serves as a critical data input for ITS. The OD matrix generation task aims to generate the synthetic OD matrix with the same data distribution as the real OD matrix and helps address the challenge of obtaining OD matrix data such as privacy concerns, data scarcity, and high data collection costs. As urban areas have expanded significantly, the OD matrix with more nodes and higher spatial resolution provide fine-scale data inputs for modeling traffic flow. Developing generative models capable of producing high-spatial-resolution OD matrix is crucial for ITS research and applications.

This paper introduces OD-PGGAN, a generative model for high-spatial-resolution OD matrix generation tasks. The proposed method employs a progressive training strategy, incorporates multi-scale generators and discriminators, and utilizes a geography-based upsampling and downsampling algorithm. Our results show that OD-PGGAN can effectively generate high-spatial-resolution OD matrix, specifically with a size of 1024×1024 . The generated synthetic OD matrix exhibits the same statistical distribution as the real OD matrix dataset and outperforms classic models, such as the gravity and radiation models.

Despite OD-PGGAN demonstrating the potential of GAN-based models for large-scale OD matrix generation tasks, there are several limitations that need to be addressed in future research. Firstly, GAN-based models typically require a substantial amount of

data to accurately capture the complex statistical properties of OD matrix. However, in real-world scenarios, OD matrix data may be limited due to privacy concerns, data scarcity, or high acquisition costs. To address this issue, this study introduces a sampling approach to construct the training dataset, ensuring sufficient data for model learning. While this method provides a viable solution, it also introduces sampling biases. Future research should explore robust data augmentation techniques to mitigate sampling biases and improve model generalizability under limited data conditions. Secondly, the regional factors such as population distribution, points of interests, urban land use, and road network density have a direct impact on traffic flows, yet they are not explicitly incorporated into the current model. Future research should consider integrating the socio-demographic factors into the generative model to improve the interpretability and accuracy of the synthetic OD matrix. Another promising future aspect is the geographical transferability of OD-PGGAN. Future studies should investigate whether OD-PGGAN trained in one city can be effectively adapted for use in other cities. Developing a generalizable generative model that integrates socio-demographic factors would help address the challenge of OD matrix estimation in cities where mobility data are scarce or unavailable. In addition, in recent years, GeoAI techniques and large language models (LLMs) have attracted significant attention in urban traffic flow research [59,60]. In future work, we plan to incorporate GeoAI and LLM-based approaches into the large-scale OD matrix generation task to further enhance model performance and scalability.

Author Contributions: Zehao Yuan, conceptualization, methodology, writing and editing; Xuanyan Chen, review and editing; Biyu Chen, review; Yubo Luo, review; Yu Zhang, review; Wenxin Teng, data collection and model solution; Chao Zhang, data collection and model solution. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Both the results and source codes can be requested from the author. The mobile phone signaling data used in this study are not accessible; however, the processed OD matrix data are available from the author upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. An, S.; Lee, B.; Shin, D. A survey of intelligent transportation systems. In Proceedings of the 2011 Third International Conference on Computational Intelligence, Communication Systems and Networks, Bali, Indonesia, 26–28 July 2011; pp. 332–337.
2. Dimitrakopoulos, G.; Demestichas, P. Intelligent transportation systems. *IEEE Veh. Technol. Mag.* **2010**, *5*, 77–84. [\[CrossRef\]](#)
3. Pavlyuk, D. Spatiotemporal Big Data Challenges for Traffic Flow Analysis. In *Reliability and Statistics in Transportation and Communication: Selected Papers from the 17th International Conference on Reliability and Statistics in Transportation and Communication, RelStat'17, Riga, Latvia, 18–21 October 2017*; Springer: Cham, Switzerland, 2018; pp. 232–240.
4. de Montjoye, Y.; Hidalgo, C.A.; Verleysen, M.; Blondel, V.D. Unique in the Crowd: The privacy bounds of human mobility. *Sci. Rep.* **2013**, *3*, 1376. [\[CrossRef\]](#)
5. Cottrill, C.D. MaaS surveillance: Privacy considerations in mobility as a service. *Transp. Res. Part A Policy Pract.* **2020**, *131*, 50–57. [\[CrossRef\]](#)
6. Kapp, A.; Hansmeyer, J.; Mihaljević, C.H. Generative models for synthetic urban mobility data: A systematic literature review. *ACM Comput. Surv.* **2023**, *56*, 1–37. [\[CrossRef\]](#)
7. Züfle, A.; Pfoser, D.; Wenk, C.; Crooks, A.; Kavak, H.; Anderson, T.; Kim, J.; Holt, N.; Diantonio, A. In Silico Human Mobility Data Science: Leveraging Massive Simulated Mobility Data (Vision Paper). *ACM Trans. Spat. Algorithms Syst.* **2024**, *10*, 1–27. [\[CrossRef\]](#)
8. Rong, C.; Ding, J.; Liu, Y.; Li, Y. A large-scale benchmark dataset for commuting origin-destination matrix generation. *arXiv* **2024**, arXiv:2407.15823.
9. Endres, M.; Mannarapotta Venugopal, A.; Tran, T.S. Synthetic data generation: A comparative study. In Proceedings of the 26th International Database Engineered Applications Symposium (IDEAS'22), Budapest, Hungary, 22–24 August 2022; pp. 94–102.

10. Jordon, J.; Szpruch, L.; Houssiau, F.; Bottarelli, M.; Cherubin, G.; Maple, C.; Cohen, S.N.; Weller, A. Synthetic Data—What, why and how? *arXiv* **2022**, arXiv:2205.03257.
11. Raghunathan, T.E. Synthetic data. *Annu. Rev. Stat. Its Appl.* **2021**, *8*, 129–140. [[CrossRef](#)]
12. Rong, C.; Ding, J.; Li, Y. An interdisciplinary survey on origin-destination flows modeling: Theory and techniques. *ACM Comput. Surv.* **2024**, *57*, 4. [[CrossRef](#)]
13. Rong, C.; Wang, H.; Li, Y. Origin-Destination Network Generation via Gravity-Guided GAN. *arXiv* **2023**, arXiv:2306.03390.
14. Saadi, I.; Mustafa, A.; Teller, J.; Cools, M. A bi-level Random Forest based approach for estimating OD matrices: Preliminary results from the Belgium National Household Travel Survey. *Transp. Res. Procedia* **2017**, *25*, 2566–2573. [[CrossRef](#)]
15. Hu, S.; Madanat, S.M.; Krogmeier, J.V.; Peeta, S. Estimation of dynamic assignment matrices and OD demands using adaptive Kalman filtering. *J. Intell. Transp. Syst.* **2001**, *6*, 281–300. [[CrossRef](#)]
16. Feng, J.; Li, Y.; Lin, Z.; Rong, C.; Sun, F.; Guo, D.; Jin, D. Context-aware spatial-temporal neural network for citywide crowd flow prediction via modeling long-range spatial dependency. *ACM Trans. Knowl. Discov. Data (TKDD)* **2021**, *16*, 49. [[CrossRef](#)]
17. Zhang, Y.; Tu, P.; Zhao, Z.; Chen, X. Incorporating prior knowledge of collision risk into deep learning networks for ship trajectory prediction in the maritime Internet of Things industry. *Eng. Appl. Artif. Intell.* **2025**, *146*, 110311. [[CrossRef](#)]
18. Hörli, S.; Becker, F.; Axhausen, K.W. Simulation of price, customer behaviour and system impact for a cost-covering automated taxi system in Zurich. *Transp. Res. Part C Emerg. Technol.* **2021**, *123*, 102974. [[CrossRef](#)]
19. Zeng, J.; Zhang, G.; Rong, C.; Ding, J.; Yuan, J.; Li, Y. Causal learning empowered OD prediction for urban planning. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22), Atlanta, GA, USA, 17–21 October 2022; pp. 2455–2464.
20. Alqahtani, H.; Kavakli-Thorne, M.; Kumar, G. Applications of generative adversarial networks (gans): An updated review. *Arch. Comput. Methods Eng.* **2021**, *28*, 525–552. [[CrossRef](#)]
21. Lin, H.; Liu, Y.; Li, S.; Qu, X. How generative adversarial networks promote the development of intelligent transportation systems: A survey. *IEEE/CAA J. Autom. Sin.* **2023**, *10*, 1781–1796. [[CrossRef](#)]
22. Figueira, A.; Vaz, B. Survey on synthetic data generation, evaluation methods and GANs. *Mathematics* **2022**, *10*, 2733. [[CrossRef](#)]
23. Mauro, G.; Luca, M.; Longa, A.; Lepri, B.; Pappalardo, L. Generating mobility networks with generative adversarial networks. *EPJ Data Sci.* **2022**, *11*, 58. [[CrossRef](#)]
24. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–9.
25. Nash, J. Two-person cooperative games. *Econom. J. Econom. Soc.* **1953**, *21*, 128–140. [[CrossRef](#)]
26. Ratliff, L.J.; Burden, S.A.; Sastry, S.S. Characterization and computation of local Nash equilibria in continuous games. In Proceedings of the 2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 2–4 October 2013; pp. 917–924.
27. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
28. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
29. Odena, A. Semi-supervised learning with generative adversarial networks. *arXiv* **2016**, arXiv:1606.01583.
30. Arjovsky, M.; Chintala, S.; Bottou, L.E.O. Wasserstein generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning, PMLR 70, Sydney, Australia, 6–11 August 2017; pp. 214–223.
31. Wang, H.; Wang, J.; Wang, J.; Zhao, M.; Zhang, W.; Zhang, F.; Xie, X.; Guo, M. GraphGAN: Graph Representation Learning with Generative Adversarial Nets. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 2508–2515.
32. Wang, Z.; Zheng, H.; He, P.; Chen, W.; Zhou, M. Diffusion-gan: Training gans with diffusion. *arXiv* **2022**, arXiv:2206.02262.
33. Ahmad, Z.; Jaffri, Z.U.A.; Chen, M.; Bao, S. Understanding GANs: Fundamentals, variants, training challenges, applications, and open problems. *Multimed. Tools Appl.* **2024**, 1–77. [[CrossRef](#)]
34. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv* **2017**, arXiv:1710.10196.
35. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
36. Brock, A.; Donahue, J.; Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv* **2018**, arXiv:1809.11096.
37. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the 36th International Conference on Machine Learning, PMLR 97, Long Beach, CA, USA, 9–15 June 2019; pp. 7354–7363.

38. Zhang, B.; Gu, S.; Zhang, B.; Bao, J.; Chen, D.; Wen, F.; Wang, Y.; Guo, B. StyleSwin: Transformer-based GAN for High-resolution Image Generation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11294–11304.
39. Kang, M.; Zhu, J.; Zhang, R.; Park, J.; Shechtman, E.; Paris, S.; Park, T. Scaling up gans for text-to-image synthesis. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 10124–10134.
40. Showrov, A.A.; Aziz, M.T.; Nabil, H.R.; Jim, J.R.; Kabir, M.M.; Mridha, M.F.; Asai, N.; Shin, J. Generative adversarial networks (GANs) in medical imaging: Advancements, applications, and challenges. *IEEE Access* **2024**, *12*, 35728–35753.
41. Oubara, A.; Wu, F.; Maleki, R.; Ma, B.; Amamra, A.; Yang, G. Enhancing adversarial learning-based change detection in imbalanced datasets using artificial image generation and attention mechanism. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 125. [\[CrossRef\]](#)
42. Šidlauskas, A.; Kriščiūnas, A.; Čalnerytė, D. Continuous Satellite Image Generation from Standard Layer Maps Using Conditional Generative Adversarial Networks. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 448. [\[CrossRef\]](#)
43. Chakraborty, T.; Ujjwal Reddy, K.S.; Naik, S.M.; Panja, M.; Manvitha, B. Ten years of generative adversarial nets (GANs): A survey of the state-of-the-art. *Mach. Learn. Sci. Technol.* **2024**, *5*, 11001. [\[CrossRef\]](#)
44. Chu, C.; Zhang, H.; Wang, P.; Lu, F. Simulating human mobility with a trajectory generation framework based on diffusion model. *Int. J. Geogr. Inf. Sci.* **2024**, *38*, 847–878. [\[CrossRef\]](#)
45. Zhang, X.; Li, N. An Activity Space-based Gravity Model for Intracity Human Mobility Flows. *Sustain. Cities Soc.* **2023**, *101*, 105073. [\[CrossRef\]](#)
46. Alis, C.; Legara, E.F.; Monterola, C. Generalized radiation model for human migration. *Sci. Rep.* **2021**, *11*, 22707. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Gu, H.; Shen, J.; Chu, J. Understanding Intercity Mobility Patterns in Rapidly Urbanizing China, 2015–2019: Evidence from Longitudinal Poisson Gravity Modeling. *Ann. Am. Assoc. Geogr.* **2023**, *113*, 307–330. [\[CrossRef\]](#)
48. Kotsubo, M.; Nakaya, T. Kernel-based formulation of intervening opportunities for spatial interaction modelling. *Sci. Rep.* **2021**, *11*, 950. [\[CrossRef\]](#)
49. Yao, X.; Gao, Y.; Zhu, D.; Manley, E.; Wang, J.; Liu, Y. Spatial origin-destination flow imputation using graph convolutional networks. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 7474–7484. [\[CrossRef\]](#)
50. Rong, C.; Li, T.; Feng, J.; Li, Y. Inferring origin-destination flows from population distribution. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 603–613. [\[CrossRef\]](#)
51. Jiang, W.; Ma, Z.; Koutsopoulos, H.N. Deep learning for short-term origin–destination passenger flow prediction under partial observability in urban railway systems. *Neural Comput. Appl.* **2022**, *34*, 4813–4830. [\[CrossRef\]](#)
52. Li, C.; Zheng, L.; Jia, N. Network-wide ride-sourcing passenger demand origin-destination matrix prediction with a generative adversarial network. *Transp. A Transp. Sci.* **2024**, *20*, 2109774. [\[CrossRef\]](#)
53. Rong, C.; Ding, J.; Liu, Z.; Li, Y. Complexity-aware large scale origin-destination network generation via diffusion model. *arXiv* **2023**, arXiv:2306.04873.
54. Youssef, A. *Image Downsampling and Upsampling Methods*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 1999.
55. Zipf, G.K. The P_1 P_2 /D Hypothesis: On the Intercity Movement of Persons. *Am. Sociol. Rev.* **1946**, *11*, 677–686. [\[CrossRef\]](#)
56. Simini, F.; González, M.C.; Maritan, A.; Barabási, A. A universal model for mobility and migration patterns. *Nature* **2012**, *484*, 96–100. [\[CrossRef\]](#)
57. Sorensen, T. *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*; I kommission hos Ejnar Munksgaard: Copenhagen, Denmark, 1948.
58. Yang, Y.; Herrera, C.; Eagle, N.; Gonzalez, M.C. Limits of predictability in commuting flows in the absence of data for calibration. *Sci. Rep.* **2014**, *4*, 5662. [\[CrossRef\]](#)
59. Zhou, Z.; Ding, J.; Liu, Y.; Jin, D.; Li, Y. Towards generative modeling of urban flow through knowledge-enhanced denoising diffusion. In Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, Hamburg, Germany, 13–16 November 2023; pp. 1–12.
60. Bhandari, P.; Anastasopoulos, A.; Pfoser, D. Urban mobility assessment using LLMs. In Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems, Atlanta, GA, USA, 29 October–1 November 2024; pp. 67–79.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.