

# Deep Learning for Cross-Domain Data Fusion in Urban Computing: Taxonomy, Advances, and Outlook

Xingchen Zou<sup>a,1</sup>, Yibo Yan<sup>a,1</sup>, Xixuan Hao<sup>a</sup>, Yuehong Hu<sup>a</sup>, Haomin Wen<sup>a</sup>, Erdong Liu<sup>a</sup>,  
Junbo Zhang<sup>b</sup>, Yong Li<sup>c</sup>, Tianrui Li<sup>d</sup>, Yu Zheng<sup>b</sup>, Yuxuan Liang<sup>a,\*</sup>

<sup>a</sup>The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

<sup>b</sup>JD Technology & JD Intelligent Cities Research, Beijing, China

<sup>c</sup>Tsinghua University, Beijing, China

<sup>d</sup>Southwest Jiaotong University, Chengdu, China

## Abstract

As cities continue to burgeon, *Urban Computing* emerges as a pivotal discipline for sustainable development by harnessing the power of cross-domain data fusion from diverse sources (e.g., geographical, traffic, social media, and environmental data) and modalities (e.g., spatio-temporal, visual, and textual modalities). Recently, we are witnessing a rising trend that utilizes various deep-learning methods to facilitate cross-domain data fusion in smart cities. To this end, we propose the first survey that systematically reviews the latest advancements in deep learning-based data fusion methods tailored for urban computing. Specifically, we first delve into data perspective to comprehend the role of each modality and data source. Secondly, we classify the methodology into four primary categories: *feature-based*, *alignment-based*, *contrast-based*, and *generation-based* fusion methods. Thirdly, we further categorize multi-modal urban applications into seven types: *urban planning*, *transportation*, *economy*, *public safety*, *society*, *environment*, and *energy*. Compared with previous surveys, we focus more on the synergy of deep learning methods with urban computing applications. Furthermore, we shed light on the interplay between *Large Language Models (LLMs)* and urban computing, postulating future research directions that could revolutionize the field. We firmly believe that the taxonomy, progress, and prospects delineated in our survey stand poised to significantly enrich the research community. The summary of the comprehensive and up-to-date paper list can be found at <https://github.com/yoshall/Awesome-Multimodal-Urban-Computing>.

**Keywords:** Urban Computing, Data fusion, Deep learning, Multi-modal data, Large language models, Sustainable development

## 1. Introduction

Cities, indispensable components of modern civilization, have undergone transformative trajectories propelled by human advancements in cultural, financial, political, and technological domains [41, 321, 390, 300]. Despite their pivotal role in societal progress, the unprecedented surge in global urbanization since the 19th century has precipitated formidable sustainability challenges including energy consumption [281, 303], environmental pollution [161, 135, 274], socio-economic disparities [330, 150], and urban traffic issues [140, 150, 212, 167]. In the 21st century, the profound strides achieved in machine learning and spatio-temporal data mining have manifested in myriad successful applications across diverse domains, such as finance [109, 215, 19], biology [194, 257], and healthcare [251, 238]. This surge in technological prowess has incited a notable shift in research focus, with scholars now directing their attention toward harnessing these advancements to optimize the intricate facets of urban planning, operations, management, etc. A pivotal contribution to this evolving discourse is the pioneering work [391], which encapsulated and elucidated

these endeavors by introducing the concept of **Urban Computing**. This paradigm leverages sensing technologies and expansive computing infrastructure to scrutinize voluminous data emanating from urban spaces. The fundamental objective is to gain profound insights into the dynamics of cities, thereby addressing challenges such as traffic congestion [119, 375, 218], energy consumption [128, 98, 223], and air quality pollution [166, 335, 334].

Urban computing necessitates the integration of extensive and diverse datasets sourced from various sources and modalities [391, 390, 67], also referred to as **Cross-Domain Data Fusion**, which arises from the recognition that relying solely on a singular data source or modality may prove inadequate for the holistic implementation of urban tasks. For example, in the realm of traffic prediction, it becomes imperative to assimilate meteorological forecast data with geographical information. This involves taking into account the congestion induced by rainfall and the influence of school and business hours on traffic flow during peak periods. In the context of urban planning, one must combine population density and economic activity data. This includes evaluating factors such as population density and income levels when devising plans for new commercial districts to ascertain their viability. Furthermore, in the field of public safety management, integrating crime data

\*Y. Liang is the corresponding author. Email: yuxiang@outlook.com

<sup>1</sup>These authors contributed equally to this work.

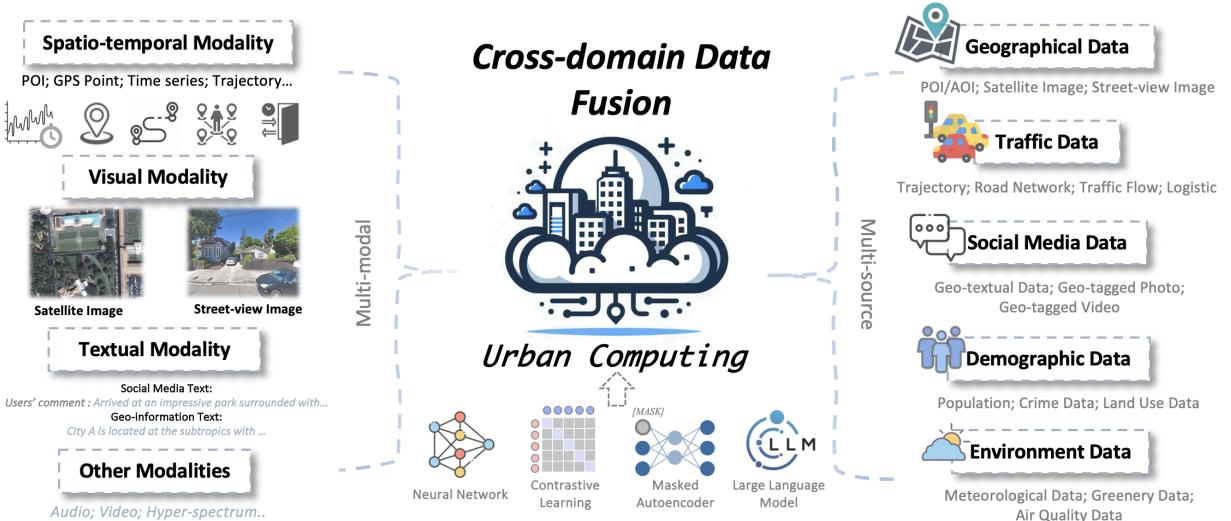


Figure 1: A sketch of cross-domain urban computing. **Left:** It involves the integration of urban data from diverse modalities, including spatio-temporal, visual, textual, and other modalities, through the process of data fusion. **Right:** Generally, these urban data derive from multiple sources, such as geographical data, transportation, social media, demography, and environment.

with socio-economic data is crucial. This entails considering unemployment rates and education levels in high-crime areas when strategizing police deployment plans. In recent years, a growing number of studies in urban computing are expanding cross-domain data fusion to encompass diverse sources like sensors [121], satellites [186], social media [380], and citizen-generated data [383]. Additionally, there is a trend towards introducing new data modalities, including text (e.g., social media posts [383] and geographic information [110, 43]) and images (e.g., satellite [186, 309, 154] and street-view images [186, 154, 112]). Figure 1 depicts the cross-domain data fusion in urban computing from the views of data modality and source.

Prior research, exemplified by Zheng et al. [391], emphasized the critical role of cross-domain data fusion in amalgamating information from multiple sources. With the emergence of data fusion studies in urban computing, Zheng [390] classified the related fusion methodologies into three types: stage-based, feature level-based, and semantic-based data fusion. Furthermore, within the purview of semantic-based methods, a more intricate taxonomy emerges, delineating four subtypes: multi-view learning-based, similarity-based, probabilistic dependency-based, and transfer learning-based methods.

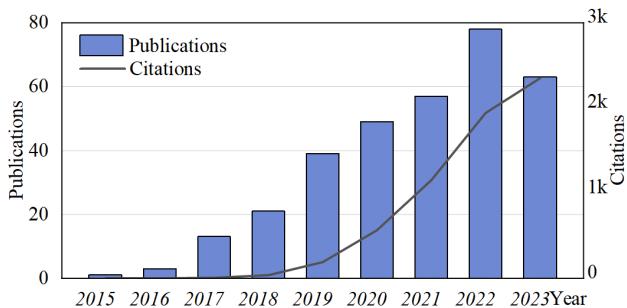


Figure 2: Times cited and publications over time for Deep Learning in Urban Computing on prestigious venues (Source: Web of Science)

Over the past decade, the emergence of Deep Learning (DL) has asserted dominance in processing spatio-temporal data in

urban computing [282, 125]. In contrast to traditional machine learning methods, these models present distinct superiority, characterized by larger model capacity, automated feature extraction capabilities, and inherent compatibility with cross-domain data fusion. As Figure 2 illustrates, there has been a significant uptick in both the number of papers published and the citations received for research related to deep learning in urban computing since 2015. The paradigm shift derived from deep learning renders previous surveys, especially [390], on urban data fusion somewhat obsolete, as traditional taxonomy may not aptly capture the nuances and differences among these advanced methodologies. In light of this issue, our survey is dedicated to bridging this gap and provides a contemporary perspective by offering a comprehensive and updated taxonomy that aligns with the era of deep learning. Through a thorough examination of deep learning-based cross-domain data fusion methods, we seek to establish a robust foundation for understanding and navigating the landscape of urban computing.

Specifically, we commence by presenting a novel taxonomy that classifies existing urban data sources into five distinct types, while concurrently categorizing fusion methods into four types. This systematic classification offers valuable insights into the intricate integration of diverse modalities within urban computing. Secondly, we systematically categorize widely used datasets and outline common application scenarios for urban computing fusion models. Thirdly, inspired by the breakthrough of Large Language Models (LLMs) in financial time series [306, 345], medicine [262, 244], law [51], and climate [236, 146], we further summarize inspiring research that integrates LLMs into urban computing, which complements the basic taxonomy of cross-domain data fusion in urban computing.

**Related Surveys.** Several surveys have recently explored the application of deep learning-based data fusion across various domains. Notably, Zheng [390] conducted a comprehensive investigation into the methodologies proposed for cross-domain big data fusion prior to 2015. Their research reveals

that machine learning-based data fusion was the de-facto dominant approach in urban computing during that period, albeit with challenges in comprehending cross-modal relations. Subsequently, with the remarkable success of deep learning models such as recurrent neural networks (RNN) and convolutional neural networks (CNN) in representation learning, Wang et al. [282] conducted an exhaustive review on the utilization of deep learning for spatio-temporal data mining, with a special focus on the fusion of multi-source spatio-temporal data. Liu et al. [177] further provided a summary of deep learning-based data fusion methodologies for urban big data fusion before 2020. Though these surveys shed light on the advancements in deep learning to their respective fields, they primarily concentrated on specific aspects and do not extensively cover cross-modal data fusion or the latest paradigms, including contrast-based and generation-based approaches which have gained significant popularity since 2021. In other words, none of them have provided a comprehensive and up-to-date taxonomy framework for data fusion methodologies in the context of urban computing.

Furthermore, existing surveys lack a specific focus on the data and application perspectives of these DL models within the field of urban computing. For instance, Gao et al. [76] summarized the fusion models for spatio-temporal data based on generative adversarial networks (GAN), while Deldari et al. [56] concentrated on self-supervised representation learning on the fusion of multi-modal data in the general domain. Besides, there are a couple of surveys investigating the deep learning-based data fusion in specific applications (i.e., a subdomain) of urban computing, such as crowd flow prediction [315], intelligent transportation [348], and social event detection [2]. These facts highlight the need for a new survey to serve as a guide for future endeavors in data fusion in urban computing concepts.

To this end, this paper aims to provide a *comprehensive* and *up-to-date* review of deep learning-based data fusion methodologies, explicitly tailored for cross-domain data fusion in urban computing. Our intention is not only documenting the latest advancements but also illuminating available resources and practical applications, and identifying potential research directions. Table 1 succinctly summarizes the key differences between our survey and other relevant surveys in the field. Through our exploration, we seek to provide a valuable resource for researchers and stakeholders, fostering an enhanced understanding of the intricacies surrounding the integration of diverse urban data modalities via deep learning approaches.

**Our Contributions.** Compared to previous surveys, the contributions of our survey can be summarized as follows:

- **Comprehensive and Up-to-Date Survey.** To the best of our knowledge, this is the first comprehensive survey that systematically reviews studies on deep learning techniques for cross-domain data fusion models in urban computing. We firmly believe that the taxonomy, progress, and prospects introduced in this paper can significantly promote the development of this field.
- **Novel and Structured Taxonomy.** We present a novel taxonomy, which organizes current research efforts from three perspectives: i) data sources, which mainly include spatio-

temporal, visual, and textual modalities; ii) fusion methods, consisting of feature-based, alignment-based, contrast-based, and generation-based fusion methods; iii) applications, spanning diverse domains such as urban planning, transportation, economy, public safety, society, environment, and energy.

- **Extensive Dataset Compilation.** Our survey thoroughly compiles and categorizes popular datasets in urban computing, taking into consideration their sources, temporal coverage, and spatial distribution characteristics. Additionally, we outline the prevailing common application scenarios for urban computing fusion models, examining their practical contributions and acknowledging their limitations in a series of downstream applications, respectively.
- **Future Research Outlook.** We endeavor to identify and provide detailed explanations of several promising directions for future research, covering various aspects, including data privacy protection, the establishment of open benchmarks, the diversification of applications, and the optimization of efficiency. Moreover, capitalizing on the progress of LLMs (e.g., GPT-4 [211] and Sora [22]) and their notable benefits in fusing multi-modal and multi-source data, we further investigate their innovative applications and propose potential approaches for urban computing.

**Organization.** The structure of the remaining sections of this paper is as follows: in Section 2, we present the overall taxonomy of deep-learning-based data fusion methods in urban computing, providing a broad overview of the field before delving into the specific perspectives and intricacies. Section 3 offers a comprehensive and detailed overview of the data utilized in urban computing, covering various modalities and sources. In Section 4, we elaborate on the fusion methods employed in urban computing, discussing their approaches and techniques. Section 5 encapsulates the extensive applications we have compiled, highlighting the practical implementations and contributions of data fusion models in urban computing. Section 7 outlines the challenges and promising avenues for future research in this domain, identifying potential areas of improvement and exploration. Finally, we conclude our paper in Section 8.

## 2. Taxonomy

This section provides a taxonomy of deep learning for multi-source and multi-modal data fusion in urban computing. As shown in Figure 2, our survey is structured along three dimensions: data in cross-domain fusion in urban computing, modality fusion methods, and applications based on data fusion. A detailed synopsis of the related works can be found in Table 1.

In Section 3, from the data perspective, we divided the data normally utilized in urban computing into five categories according to data sources: *geographical data*, *traffic data*, *social media data*, *demographic data* and *environmental data*. Additionally, we have also classified the data from a modality perspective, including *spatio-temporal data*, *visual data*, *textual data*, and other types of data. These two categorizations allow for a systematic understanding and analysis of the different

Table 1: A thorough comparison between related surveys and ours, focusing on scopes (i.e., Specific versus Urban Computing), relevant modalities (e.g., general spatio-temporal data, image, text, and others), and primary topics of focus (i.e., data sources and modalities utilized for urban computing (Data), data fusion models and techniques (Fusion Model), application domains and downstream tasks in urban computing (Application) and LLMs for urban computing).

Survey	Year	Venue	Scope		Modality			Focus				
			Specific	Urban Computing	Spatio-temporal	Image	Text	Others	Data	Fusion Model	Application	🔥 LLM
Zheng [390]	2015	IEEE Trans. Big Data		✓	✓	✓	✓	✗	✗	✓	✗	✗
Wang et al. [282]	2020	IEEE TKDE	✓		✓	✗	✗	✗	✓	✓	✓	✗
Liu et al. [177]	2020	Information Fusion		✓	✓	✓	✓	✓	✗	✓	✗	✗
Xie et al. [315]	2020	Information Fusion	✓		✓	✗	✗	✗	✓	✓	✗	✗
Yuan and Li [348]	2021	Springer Data Sci. Eng.	✓		✓	✗	✗	✗	✓	✓	✓	✗
Afyouni et al. [2]	2022	Information Fusion	✓		✓	✓	✓	✓	✓	✗	✓	✗
Gao et al. [76]	2022	ACM TIST	✓		✓	✗	✗	✗	✗	✓	✓	✗
Deldari et al. [56]	2022	Unpublished	✓		✓	✓	✓	✓	✗	✓	✗	✗
<b>Ours</b>	2024	-		✓	✓	✓	✓	✓	✓	✓	✓	✓

types of data used in urban computing research. We further present a comprehensive overview of public datasets used in cross-domain data fusion in urban computing in Table 2.

From the fusion methodology perspective, Section 4 covers a comprehensive review of existing data fusion methods in urban computing, categorized into *feature-based*, *alignment-based*, *contrast-based*, and *generation-based fusion*. In each category, we subdivide the existing literature into several types based on the models' properties. A detailed taxonomy from a modality fusion perspective can be found in Figure 2. Additionally, in the generation-based fusion section, we also focus on the recent application of LLM for data fusion in urban computing which offers valuable insights for the research community.

In Section 5, we divide the multi-modal application in urban computing into seven categories: *urban planning*, *transportation*, *economy*, *public safety and security*, *social*, *environment* and *energy*. We explore the superiority of multi-modal data fusion methodologies in each type of downstream task.

### 3. Data Perspective

This section delves into the datasets used in cross-domain data fusion in urban computing. Based on the literature available since 2015, we categorize diverse urban data and perform a statistical analysis of their distributions. Furthermore, we discuss how each type of data is incorporated into different research and real-world scenarios.

#### 3.1. Overview

Based on the characteristics of the data and their diverse sources from various domains, this survey categorizes datasets utilized in the field of cross-domain data fusion in urban computing into six segments, including geographical data, traffic data, social network data, demographic data, environment data, and other data (i.e., data that cannot be categorized into the aforementioned types, such as healthcare data). As illustrated in Figure 1 and 2, these categories are defined as follows:

- **Geographical data** refers to geographical information of specific locations on the Earth's surface, such as coordinates (i.e., latitude and longitude). This category extends to spatial attributes, including but not limited to topography, land use, and physical features.

- **Traffic data** encompasses information related to the movement of vehicles and pedestrians, including factors like traffic flow, congestion, speed, and road conditions.

- **Social media data** comprises user-generated content from online platforms, encompassing geo-tagged text, images, and videos, offering insights into user behaviors, sentiment, and emerging trends.

- **Demographic data** involves statistical information about populations, including characteristics such as age, gender, ethnicity, income, and education.

- **Environment data** incorporates information about the natural world, covering aspects like climate, air quality, biodiversity, and pollution levels.

Based on the aforementioned categorization, Table 2 summarized open-sourced datasets commonly used for cross-domain data fusion in urban computing. Figure 4 presents the distribution of dataset modalities across all investigated papers in this survey. From the pie chart, it is evident that geographical and transportation data are the most crucial datasets in urban computing, with a majority of papers (approaching 70%) opting for them as a primary modality. Following closely is social media data, which plays a significant role in urban computing, often combined with other modal datasets to address the complex and dynamic challenges of urban environments. Demographic data and environment data are also common datasets, however, only around 11% of studies choose to incorporate these two types of data in their research.

Furthermore, we also investigate the geographic distribution of the datasets across various cities and countries. As depicted in Figure 5, the bar chart indicates the frequency of dataset utilization in different cities (i.e., bars) and countries (i.e., colors). Distinctive colors of the bars denote different countries, indicating the popularity of the data focus in certain countries. Notably, datasets from Beijing and New York emerged as extensively utilized, followed closely by cities such as Chicago, Singapore, and Shanghai. Overall, the majority of datasets in the domain of cross-domain data fusion in urban computing originate from China and the United States.

Table 2: Taxonomy and summary of open-sourced dataset used for cross-domain data fusion in urban computing.

Category	Content	Format	Dataset	Link	Reference
Geographical Data	Satellite Image	Image	ArcGIS	<a href="https://developers.arcgis.com">https://developers.arcgis.com</a>	[186]
			PlanetScope	<a href="https://developers.planet.com/docs/data/planetscope/">https://developers.planet.com/docs/data/planetscope/</a>	[154]
			Google Earth	<a href="https://developers.google.com/maps/documentation/">https://developers.google.com/maps/documentation/</a>	[116]
			OpenStreetMap	<a href="https://www.openstreetmap.org/">https://www.openstreetmap.org/</a>	[337]
	Street-View Image	Image	Baidu Maps	<a href="https://lbsyun.baidu.com">https://lbsyun.baidu.com</a>	[324, 313]
			Baidu Map	<a href="https://lbsyun.baidu.com">https://lbsyun.baidu.com</a>	[186, 124]
			Google Street	<a href="https://developers.google.com/maps/">https://developers.google.com/maps/</a>	[186, 4]
	POIs	Point Vector	Tencent Map	<a href="https://lbs.qq.com/tool/streetview/index.html">https://lbs.qq.com/tool/streetview/index.html</a>	[112]
			Tencent Map Service	<a href="https://lbs.qq.com/getPoint/">https://lbs.qq.com/getPoint/</a>	[309, 235]
			WeChat POIs	<a href="https://open.weixin.qq.com">https://open.weixin.qq.com</a>	[277]
Traffic Data	Traffic Trajectory	Spatio-temporal Trajectory	Baidu Map POIs	<a href="https://lbsyun.baidu.com">https://lbsyun.baidu.com</a>	[154, 172, 175, 110, 313]
			NYC Open POIs	<a href="https://opendata.cityofnewyork.us/">https://opendata.cityofnewyork.us/</a>	[170, 272, 20, 366, 288]
			Foursquare	<a href="https://developer.foursquare.com/docs/checkins/checkins">https://developer.foursquare.com/docs/checkins/checkins</a>	[20, 381, 13, 42, 107, 116]
			Wikipedia POIs	<a href="https://www.wikipedia.org">https://www.wikipedia.org</a>	[386]
			AMap Service	<a href="https://lbs.amap.com">https://lbs.amap.com</a>	[10]
			Yelp POIs	<a href="https://www.yelp.com/developers">https://www.yelp.com/developers</a>	[13, 380, 383]
			Dianping POIs	<a href="https://api.dianping.com/">https://api.dianping.com/</a>	[33, 63]
			Weibo POIs	<a href="https://open.weibo.com/wiki/API">https://open.weibo.com/wiki/API</a>	[33, 134, 77]
			Flickr POIs	<a href="https://www.flickr.com/services/developer/api/">https://www.flickr.com/services/developer/api/</a>	[99]
			Bing Map POIs	<a href="https://www.bingmapsportal.com">https://www.bingmapsportal.com</a>	[37]
Social Media Data	Text	Text	Shenzhou UCar	<a href="https://bit.ly/2MG47xz">https://bit.ly/2MG47xz</a>	[93]
			Chicago Transportation	<a href="https://data.cityofchicago.org/">https://data.cityofchicago.org/</a>	[272, 288, 116]
			VED	<a href="https://github.com/gsobh/VED">https://github.com/gsobh/VED</a>	[209, 372]
			Taxi Shenzhen	<a href="https://github.com/cbdog94/STL">https://github.com/cbdog94/STL</a>	[113, 302]
			NYC Open Taxi Data	<a href="https://opendata.cityofnewyork.us/how-to/">https://opendata.cityofnewyork.us/how-to/</a>	[368, 366]
			GeoLife	<a href="http://urban-computing.com/index-893.htm">http://urban-computing.com/index-893.htm</a>	[96, 398, 400, 394, 347]
			T-Drive Taxi	<a href="http://urban-computing.com/index-58.htm">http://urban-computing.com/index-58.htm</a>	[350, 351, 217, 191]
			DiDi Traffic	<a href="https://outreach.didichuxing.com/research/opendata/">https://outreach.didichuxing.com/research/opendata/</a>	[349, 188, 228, 328, 261]
			Xiamen Taxi	<a href="https://data.mendeley.com/datasets/6xg39x9vgd/1">https://data.mendeley.com/datasets/6xg39x9vgd/1</a>	[342, 40, 124, 39]
			Grab-Posisi	<a href="https://goo.su/W3yD5m">https://goo.su/W3yD5m</a>	[337, 339]
Demographic Data	Taffic Flow	Spatio-temporal Graph	California-PEMS	<a href="http://pems.dot.ca.gov">http://pems.dot.ca.gov</a>	[9, 254]
			METR-LA	<a href="https://www.metro.net">https://www.metro.net</a>	[143, 171]
			Large-ST	<a href="https://github.com/liuxu77/LargeST">https://github.com/liuxu77/LargeST</a>	[182]
			MobileBJ	<a href="https://github.com/FIBLAB/DeepSTN/issues/4">https://github.com/FIBLAB/DeepSTN/issues/4</a>	[170, 134, 33]
			TaxiBJ	<a href="https://goo.su/aQyjTAz">https://goo.su/aQyjTAz</a>	[164, 11, 226, 120, 368, 74]
	Road Network	Spatial Graph	BikeNYC	<a href="https://citibikenyc.com/">https://citibikenyc.com/</a>	[170, 11, 226, 120]
			OpenStreetMap	<a href="https://www.openstreetmap.org">https://www.openstreetmap.org</a>	[339, 13, 188, 349, 84]
			US Census Bureau	<a href="https://www.census.gov/data.html">https://www.census.gov/data.html</a>	[366]
			LaDe	<a href="https://cainiao.techai.github.io/LaDe-website/">https://cainiao.techai.github.io/LaDe-website/</a>	[305]
			JD Logistics	<a href="https://corporate.jd.com/ourBusiness#jdLogistics">https://corporate.jd.com/ourBusiness#jdLogistics</a>	[235]
Environment Data	Users' Info	Time Series	Twitter	<a href="https://developer.twitter.com/en/docs">https://developer.twitter.com/en/docs</a>	[20, 381, 383, 352, 270, 301, 240]
			Common Crawl	<a href="https://registry.opendata.aws/commoncrawl/">https://registry.opendata.aws/commoncrawl/</a>	[289, 283, 285, 284, 200, 184]
			Yelp Reviews	<a href="https://www.yelp.com/dataset">https://www.yelp.com/dataset</a>	[380]
			Weibo Traffic Police	<a href="http://open.weibo.com/developers/">http://open.weibo.com/developers/</a>	[380, 383]
			Geotagged Image & Video	<a href="https:// goo.su/jzaDU">https:// goo.su/jzaDU</a>	[342]
	Crime	Time Series	YFCC100M	<a href="https:// goo.su/dWPQZcd">https:// goo.su/dWPQZcd</a>	[386, 340, 99]
			NUS-WIDE	<a href="https:// qualinet.github.io/databases/video/">https:// qualinet.github.io/databases/video/</a>	[340, 338]
			GeoUGV	<a href="https:// goo.su/187">https:// goo.su/187</a>	[187]
			Jie pang User Check-in	<a href="https:// jiepang.app/">https:// jiepang.app/</a>	[74]
			Gowalla User Location	<a href="https:// konect.cc/networks/loc-gowalla_edges/">https:// konect.cc/networks/loc-gowalla_edges/</a>	[42, 352]
Air Quality	Population	Time Series	WeChat Mobility	<a href="https:// open.weixin.qq.com">https:// open.weixin.qq.com/</a>	[277]
			NYC Crime	<a href="https:// opendata.cityofnewyork.us/">https:// opendata.cityofnewyork.us/</a>	[368]
			Land Use SG	<a href="https:// www.ura.gov.sg/Corporate/Planning/Master-Plan">https:// www.ura.gov.sg/Corporate/Planning/Master-Plan</a>	[156]
			Land Use NYC	<a href="https:// goo.su/putuG">https:// goo.su/putuG</a>	[156]
			WorldPop	<a href="https:// www.worldpop.org/">https:// www.worldpop.org/</a>	[309, 154, 10]
Metereology	Greenery	Time Series	TipDM China Weather	<a href="https:// www.tipdm.org/">https:// www.tipdm.org/</a>	[178]
			DarkSky Weather	<a href="https:// support.apple.com/en-us/102594">https:// support.apple.com/en-us/102594</a>	[349]
			WeatherNY	<a href="https:// opendata.cityofnewyork.us/">https:// opendata.cityofnewyork.us/</a>	[272]
			WeatherChicago	<a href="https:// data.cityofchicago.org/">https:// data.cityofchicago.org/</a>	[272]
			Weather Underground	<a href="https:// www.wunderground.com/">https:// www.wunderground.com/</a>	[342]
			DidiSY	<a href="https:// www.didiglobal.com/">https:// www.didiglobal.com/</a>	[12]
			WD_BJ weather	<a href="https:// goo.su/DmHFHd">https:// goo.su/DmHFHd</a>	[192]
			WD_USA weather	<a href="https:// goo.su/RvhBA">https:// goo.su/RvhBA</a>	[192]
			Google Earth	<a href="https:// earth.google.com/">https:// earth.google.com/</a>	[342]
			UrbanAir	<a href="https:// goo.su/hfzNB53">https:// goo.su/hfzNB53</a>	[399, 396, 392]
			KnowAir	<a href="https:// github.com/shuowang-ai/PM2.5-GNN">https:// github.com/shuowang-ai/PM2.5-GNN</a>	[286, 346, 370, 318]

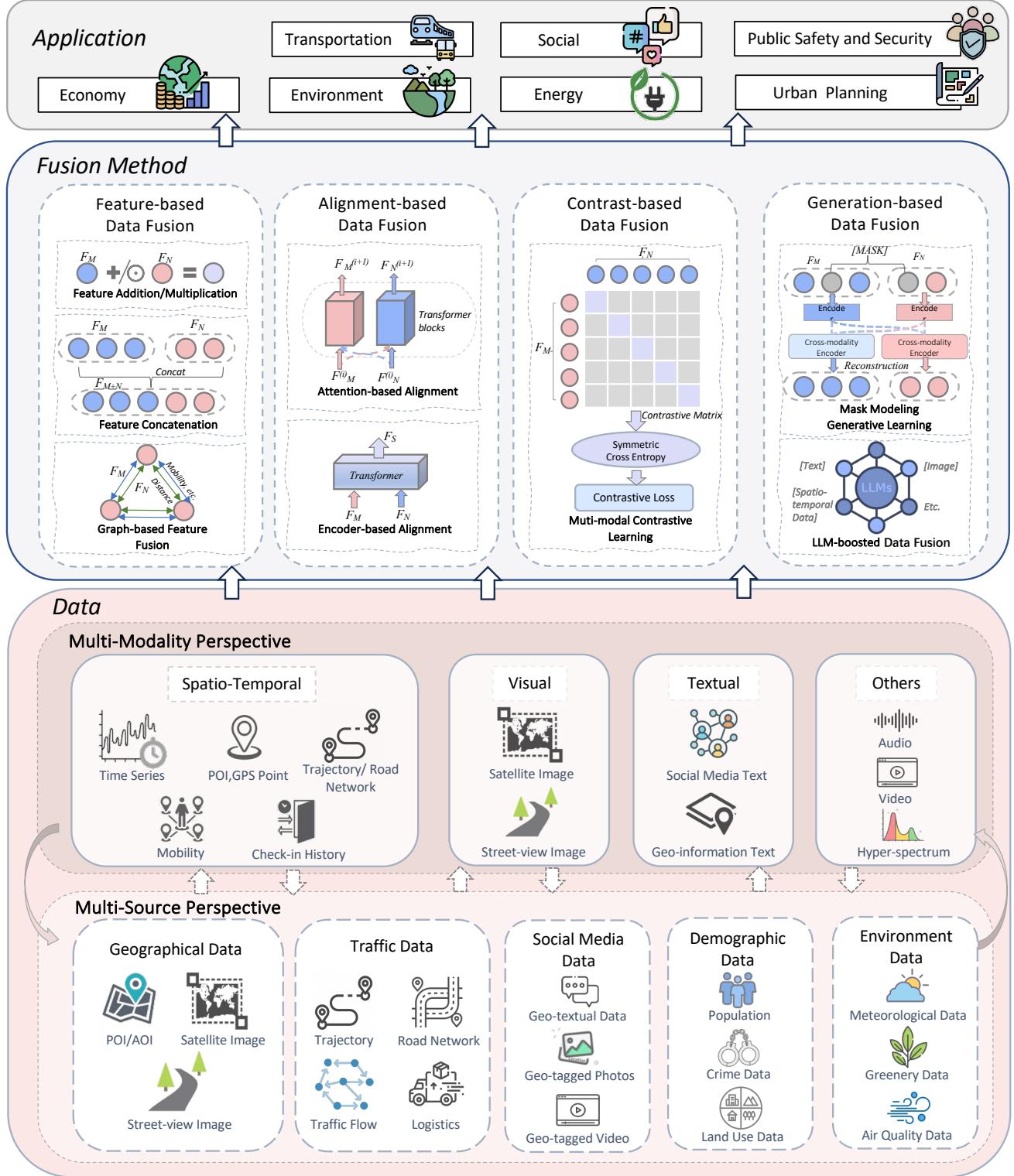


Figure 3: The taxonomy framework for deep learning-based cross-domain data fusion in urban computing in our survey. The framework is structured around three dimensions: data, fusion method, and application. Within each perspective, we categorize existing research into different categories to provide a comprehensive and well-organized review.

### 3.2. Geographical Data

Geographical data plays a crucial role in modeling spatial relationships, contributing to the enhancement of cross-domain

data fusion in urban computing by providing valuable insights into the geospatial context [21, 379, 37]. Tobler's First Law of Geography, as stated by Miller [198], indicates that "Everything

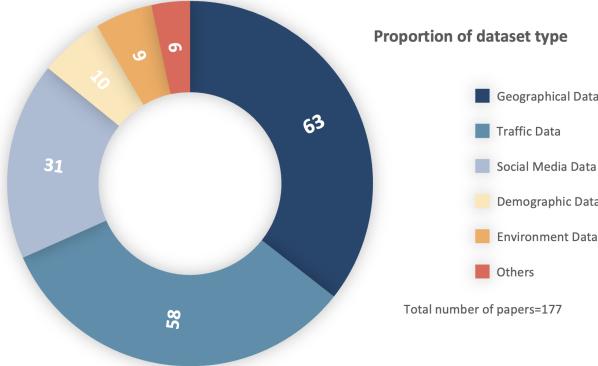


Figure 4: Proportion of dataset type among highly related papers within the scope of cross-domain data fusion in urban computing.

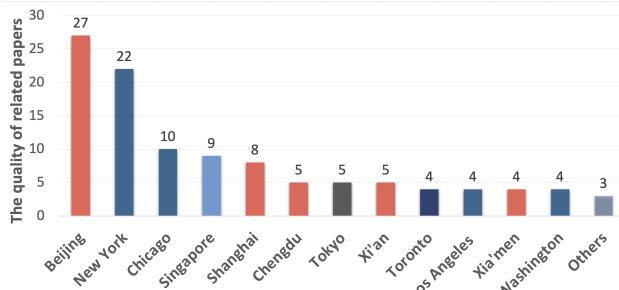


Figure 5: Distribution of dataset usage frequency from different cities (i.e., bars) and countries (i.e., colors) across highly related papers within the scope of this survey. Note that the cities with less than four paper usage are omitted in this illustration for simplicity.

is related to everything else, but near things are more related than distant things.” This emphasizes the importance of geographical data, which serve as the basis for spatial modeling-based urban computing research. Figure 6 shows the main four types of geographical data through a visualization approach.

**Points of Interest (POI) data**, as a cornerstone of cross-domain data fusion in urban computing, command significant attention and utility in the realm of geographical data [170, 272, 386, 13, 69, 31, 79, 107]. POI data refers to a collection of data representing specific locations or sites that hold significance or interest, often encompassing businesses, landmarks, or other notable entities in a given geographical area [224, 177]. Conventionally, POI data may include the location coordinates (i.e., longitude and latitude), address, categorization (e.g., restaurants, hotels, parks, etc.), address, phone number, operating hours, and so on. Recent cross-domain urban datasets may also contain user reviews, photos, and other individual information [380, 33, 63, 13]. Therefore, POI data is capable of representing the semantics of a specific area, encapsulating key locations and entities to provide a comprehensive understanding of the geographic context. For example, Bing et al. [20] collected POI data of two cities: New York City and Tokyo from the Foursquare platform and observed that POIs with higher spatial similarity values often have similar semantics. Through AMaps Service Platform in Wuhan and Shanghai, Bai et al. [10] collected approximately 2 million POI records and fused them with satellite visual embedding for socioeconomic downstream tasks. To model the spatial correlations be-

tween POIs, Fu et al. [74] developed a spatial network based on the multi-relation data among POIs, such as spatial distance and mobility connectivity.



Figure 6: Visualization of four types of geographical data collected at Central Park, New York, USA: (a) POI data and digital map; (b) street-view image; (c) satellite image.

**Satellite image data** holds significant importance in representing geo-context by providing a visual depiction of Earth’s surface, and offering advantages such as global coverage, real-time monitoring, and the ability to capture fine-scale details [26, 353, 207]. Satellite imagery is readily accessible and amenable to processing through various publicly available platforms (e.g., Google Earth [116], OpenStreetMap [337], Baidu Map [324, 313], and PlanetScope [154]), facilitating multimodal research in urban computing domain.

Moreover, **street-view image data** plays a vital role in depicting geo-contextual information by providing immersive, ground-level perspectives [18, 88, 130]. A significant amount of cross-domain urban research has collected data from public platforms such as Google Street [186, 4], Baidu Map [186, 124], and Tencent Map [113].

### 3.3. Traffic Data

Traffic data accounts for the second largest proportion in Figure 4. Different from geographical data, traffic data is generated through human activities and is directly associated with socio-economic factors, making it a distinct data type with significant implications. Traffic data finds application in diverse downstream tasks, encompassing multiple domains within urban computing, which involve spatial and temporal dimensions, as well as dynamic and static aspects. Therefore, we classify traffic data into four distinct types, taking into account their characteristics and usage scenarios: trajectory data, traffic flow data, road network data, and other miscellaneous data. Figure 7 indicates the visualization of the first three types of traffic data.

**Human trajectory data** can be conceptualized as a sequential trace produced by an object in motion within geographical spaces. This trajectory is typically represented as an array of chronologically sequenced points, delineated as  $p_1 \rightarrow p_2 \rightarrow \dots \rightarrow p_n$ . Each point in this sequence comprises a set of geospatial coordinates coupled with a corresponding timestamp, formalized as

$$\mathbf{P} = (\mathbf{x}, \mathbf{y}, t), \quad (1)$$

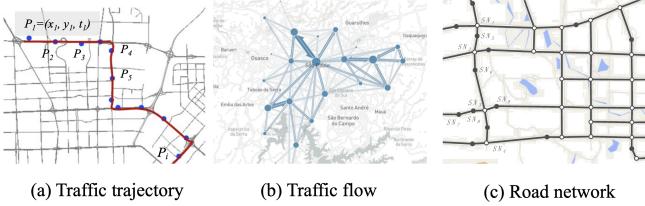


Figure 7: Visualization of three primary types of traffic data: (a) traffic trajectory; (b) traffic flow; (c) road network [307].

where  $\mathbf{x}$  and  $\mathbf{y}$  denote the spatial coordinates, typically representing longitude and latitude in a geodetic framework.  $\mathbf{t}$  represents the timestamp. This format facilitates the precise tracking and analysis of spatial dynamics over time.

Nowadays, location-based services, such as taxi and ride-hailing services, have generated enormous trajectory data. Additionally, many service providers, such as Uber and DiDi, have made their datasets publicly available to support research in this field. For instance, Zhang et al. [368] utilized taxi trip data during one month as a modality to assist in profiling urban regions. PANDA [342] demonstrated the significance of taxi trajectory in predicting road risk, which can aid in the extraction of road segments. Geng et al. [84] forecasted ride-hailing demand on large-scale ride-hailing datasets in Beijing and Shanghai.

**Traffic flow data** provides critical insights into the movement patterns of vehicles and pedestrians within urban environments, essential for understanding and optimizing city dynamics. In the realm of urban computing, three datasets — *MobileBJ*, *BikeNYC*, and *TaxiBJ* — emerge as key sources of insight for traffic mobility. Specifically, the *MobileBJ* dataset, collected from a Chinese social network, has significantly contributed to three key areas of urban computing. Firstly, it enabled the development of DeepSTN+ [170], a deep learning model for predicting urban crowd flows, which integrates spatial dependencies, POIs, and temporal data. Additionally, it supported studies on high-dimensional spatio-temporal data in urban communities, employing advanced representation learning techniques [134, 33].

The *BikeNYC* dataset, derived from New York City’s bike-sharing system in 2014, provides rich data on trip durations, station locations, and times. It focuses on the last 14 days for testing purposes and includes 9 POI types. This dataset has been instrumental in developing a model for multi-step passenger demand forecasting [11], leveraging its detailed trip data to address complex spatio-temporal challenges in urban transportation planning [120, 226, 367]. Meanwhile, *TaxiBJ* collects GPS traces from Beijing taxis, encompassing trip details, travel times, speeds, and specific pick-up and drop-off points. This dataset allowed for the development of a novel embedding strategy for urban regions [74], providing deeper insights into city dynamics and supporting sustainable urban development. It also enhanced traffic prediction accuracy by applying spatiotemporal graph neural networks and integrating self-supervised learning for improved long-term forecasting [120]. Furthermore, *TaxiBJ* demonstrated the effectiveness of a con-

trastive self-supervision method in inferring fine-grained urban flows, especially in environments with limited resources [226]. An innovative approach for computing trajectory similarities, *TrajGAT*, was introduced using *TaxiBJ*, significantly advancing the analysis of long trajectory data [328].

**Road network data** is intimately related to human daily lives, as it serves as the foundation for various services such as navigation and food delivery. The acquisition of road network data can be achieved through various methods. The earliest on-site manual surveying was labor-intensive and required a large amount of resources. With the development of remote sensing technology, which can effectively reduce costs, the proportion of road network data drawing relying on on-site collection has been decreasing. Another method to collect road network data is through UGC (User Generated Content), which collects road network information through anonymous terminal devices. The collected route can effectively help with updating road attributes and refinement of the road network. Generally, road network data can be downloaded in shapefile format from different open-sourced platforms such as Open Street Map, GRIP global roads database, DIVA-GIS, etc. For example, Yuan et al. [349] predicted travel demand and traffic flow based on taxi orders which is associated with trajectories on the road network. Zhu et al. [405] ranked region significance taking into account multi-source spatial data including trajectories on road networks. Besides, the acquisition of street-view image datasets [112, 68] also necessitated the sampling of collection points on road networks.

Other miscellaneous data include logistic data, transportation safety data, and transportation recommendation data. In the logistics field, LaDe [305] introduced the first industry-scale last-mile delivery dataset. LaDe includes detailed information about the courier trajectories and waybill information, which can support massive spatio-temporal data mining tasks [235]. Transportation safety data is also of great significance. For example, in [39], road obstacle data was used to develop RADAR, a real-time system for identifying road obstacles in urban areas during typhoon seasons. You et al. [342] used a comprehensive event dataset, including data on road accidents, fallen trees, and ponding water, to successfully develop a framework for predicting road risks in post-disaster urban settings with high accuracy. Transportation recommendations have become an integral part of our daily lives, contributing significantly to various service enhancements. For instance, Gao et al. [78] utilized this data to develop GraphTrip, a groundbreaking framework that leverages spatio-temporal graph representation learning for trip recommendations. This framework was rigorously tested using datasets from Edinburgh, Glasgow, Osaka, Toronto, and Melbourne, showcasing notable improvements in travel planning accuracy. Similarly, Guo et al. [93] and He et al. [99] employed multi-source urban data, encompassing operational data, taxi GPS trajectories, and public transportation information, to create advanced recommendation models.

### 3.4. Social Media Data

Twitter, an online social media and social networking service allows registered users to post **geo-textual data** [50]. Hence,

the Tweet data, owing to its inherent geo-tagged information, is utilized by researchers as a modality representing user social states, and it is integrated with other models in multi-modal learning. For instance, Zhao et al. [381] collected English tweets using the Twitter API in two cities, New York City and Singapore, and then annotated whether a tweet was POI-related. Finally, they can model the association between the tweet and its most semantically related POI. To study periodic human mobility patterns, Yuan et al. [352] collected geo-annotated tweets from the most recent 3,200 tweets of Twitter users, and mapped them to the corresponding city by reverse geocoding. With a similar goal of understanding urban dynamics, Miyazawa et al. [200] focused only on tweets regarding mobility and social activity; while Wang et al. [284] are more interested in tweets concerning traffic events in Chicago. Furthermore, geo-textual data such as tweets can be used to learn user preferences to support personalized maps [383]. Some work, such as [270] working on POI boundary estimation, also removed the content automatically created by other services like Twimight, Tweetbot, and so forth.

**Geo-tagged photos** are crucial in multi-modal learning as they provide spatial context, enriching the understanding of content by incorporating location information into the broader spectrum of data modalities [203, 7, 25]. The two most commonly used geo-tagged photo datasets are the Yahoo Flickr Creative Commons (YFCC) dataset [263] as well as the NUS-WIDE dataset [47]. It is noteworthy that the YFCC dataset is well-known as the largest public multimedia collection released, which consists of 100 million photos posted on Flickr with relevant meta information such as geo-location coordinates and the date taken. The NUS-WIDE dataset is a well-known web image dataset (with 269,648 images and the associated tags from Flickr) created by Lab for Media Search in the National University of Singapore. Both Zhao et al. [386] and He et al. [99] leveraged the geo-tagged photos to jointly learn a context-aware embedding for personalized tour recommendation. Yin et al. [338] derived training labels from the geo-tagged documents from the NUS-WIDE dataset, and therefore generated GPS embeddings. In the work of Yin et al. [340] where multi context-aware location representations need to be learned, the geo-tagged photos from the NUS-WIDE dataset are used for image classification evaluation, whereas those from the YFCC dataset are leveraged to learn the semantic context.

With the ubiquity of sensor-rich smartphones, acquiring continuous video frames with spatial metadata has become practical. The public **geo-tagged mobile video data** comes mainly from two mobile platforms, MediaQ [137, 197] and GeoVid [9, 85]. Lu et al. [187] collected the geo-tagged video data from the aforementioned data platforms, and proposed the GeoUGV dataset consisting of two sets, videos and their geospatial metadata. Moreover, many social media platforms (e.g., WeChat [277], Gowalla [42, 352], Baidu [172], Jiepong [74]) encompass valuable user information, serving as a fundamental modality for fusion.

### 3.5. Demographic Data

The inclusion of demographic datasets in spatio-temporal multi-modal learning is pivotal as it enhances the contextual understanding of a given human group [246, 271, 252]. The WorldPop organization plays a crucial role in global demographic research by providing high-resolution **population data**, enabling informed decision-making, and addressing various socio-economic and public health challenges worldwide [3, 189, 45, 233]. Xi et al. [309] and Li et al. [154] collected Beijing population and population density statistics from the WorldPop platform as one of the predicted socioeconomic indicators; whereas Li et al. [156] extracted those from Singapore and New York City. In terms of data post-processing, Bai et al. [10] further estimated the population density per grid cell through the Random Forest-based redistribution method.

**Crime data** holds immense significance as it serves as a critical resource for understanding patterns and factors influencing criminal activities, enabling policymakers and researchers to develop effective strategies for crime prevention and public safety [108, 95, 376]. For example, Zhang et al. [368] collected crime data from the NYC Open Data website as ground truth values to be predicted by region embeddings, and there are 40 thousand crime records during one year in New York City.

Besides, **land use data** is crucial in urban planning, providing valuable insights into the spatial distribution of human activities, and informing decision-makers to optimize land resources for various purposes [250, 35, 231]. It consists of property data (e.g., private residential property transactions), residential data (e.g., buying and renting properties), business data (e.g., renewal of business use), and so on. For instance, Li et al. [156] collected land use data of Singapore and New York City from Singapore Master Plan 2019 and NYC MapPLUTO, respectively, as a region representation evaluation benchmark.

### 3.6. Environment Data

Environment data, particularly **meteorological data**, offers essential insights into dynamic weather patterns and environmental conditions that are integral for understanding complex interactions in various urban domains [101, 8, 222, 266]. The publicly available APIs can be utilized to gather meteorological data. For example, Yuan et al. [349] utilized the Dark Sky API integrated into Apple Weather to extract diverse weather characteristics for each region and used one-hot encoding to represent categorical attributes. You et al. [342] extracted temporal contextual features including rainfall, temperature, humidity, dew point, and wind speed using Weather Underground API. Wang et al. [272] obtained temperature and sky condition data for both NYC and Chicago from NYC Open Data and Chicago Data Portal, respectively. By doing so, they were able to analyze the spatio-temporal correlations to predict the risk of traffic accidents. To precisely forecast the weather, Ma et al. [192] employed three real-world weather datasets: the WD\_BJ dataset, which was gathered from 10 ground automatic weather stations in Beijing and contains nine meteorological variables; and the WD\_ISR dataset and WD\_USA dataset, both collected from OpenWeather and providing information on four

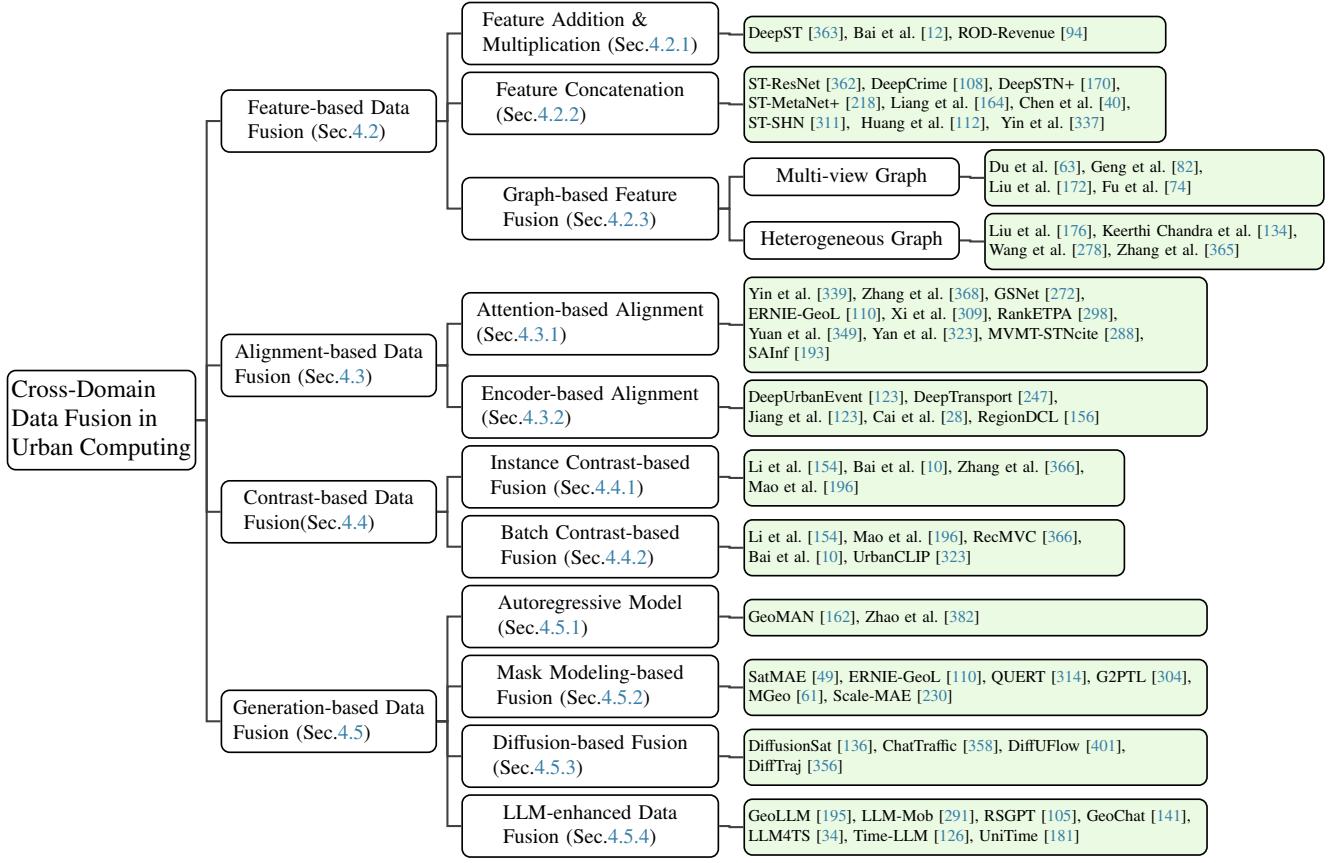


Figure 8: Taxonomy of deep learning-based cross-domain data fusion methods in urban computing.

weather conditions (namely temperature, humidity, wind speed, and atmospheric pressure) for Israel and the USA, respectively. In addition, enterprises also gather their own meteorological data, such as the DidiSY dataset which includes information on weather conditions, temperature, and wind speed specifically in Shenyang, a major city in China [12]. Besides, open-source datasets from competitions can also be utilized, such as the Chinese weather dataset from the 7th Teddy Cup Data Mining Challenge, which includes 12 different types of weather [178].

**Greenery data** holds significance in sustainable urban planning, providing valuable information about the distribution and density of vegetation, which is essential for assessing biodiversity, urban green spaces, and their impact on overall ecological health [333, 38, 208]. In particular, You et al. [342] extracted the degree of tree coverage using AlexNet [139] based on satellite imagery obtained from the Google Earth platform. **Air quality data** is of paramount importance for environmental management, facilitating the identification of sources, and enabling the development of effective strategies to mitigate the adverse impacts of air pollution on both human health and the environment [374, 260, 32]. For example, the UrbanAir system offers air quality data every hour from 2,296 stations in 302 Chinese cities, where each record consists of the concentration of six pollutants:  $NO_2$ ,  $SO_2$ ,  $O_3$ ,  $CO$ ,  $PM2.5$  and  $PM10$  [399].

## 4. Methodology Perspective

In this section, we begin by providing definitions and explanations of four types of multi-modal fusion. Then, we explore each type in detail, offering more specific categorizations and providing illustrative examples for each case. The comprehensive summary of fusion models can be found in Table 3.

### 4.1. Overview

The integration of deep learning techniques into urban computing has facilitated the development of various deep-learning-based data fusion methods. These techniques aim to leverage the inherent connections within diverse urban data streams. To better understand the differences in the underlying concepts of data fusion in these studies, we consult prior surveys on methodologies in Urban Computing [390, 177] and taxonomies of deep learning methods in other fields [282, 348, 56, 76] and categorize the lastest researches into four distinct groups based on their underlying techniques, as illustrated in Figure 8. The definitions of each category are outlined as follows:

**Feature-Based Data Fusion** is a straightforward fusion strategy, involving the combination of features from diverse sources such as sensor, visual, and textual data. The core idea is to consolidate raw or processed data features from various sources, forming comprehensive characteristics of studied urban objects. This integration is crucial for enhancing the predictive capabilities of urban models, facilitating complex analy-

Table 3: The summary of deep learning-based cross-domain data fusion models in urban computing. We denote different data source as follows: Traffic Data ; Geographical Data ; Social Media Data ; Demographical Data ; Environmental Data .

Category	Method	Data Source	Modality					Application	Institution	Year	
			General Spatio-temporal			Visual		Textual			
Time series	POI / Location	Trajectory / Road network	Mobility	ST events	Satellite image	Street-view image	Social media text	Geo-information text	Application	Institution	Year
Feature Based Data Fusion	DeepST [363]								Transportation	Microsoft	2016
	ST-ResNet [364]								Transportation	Microsoft	2018
	ST-MetaNet+ [218]								Transportation	JD Research	2020
	DeepCrime [108]								Social	JD Research	2021
	STUKG [278]								Transportation	THU	2021
	DeepSTN+ [170]								Transportation	THU	2019
	DeepTP [349]								Transportation	THU	2021
	Guo et al. [94]								Transportation	BUAA	2019
	Photo2Trip [386]								Transportation	SU/RU/UCA	2017
	ST-SHN [311]								Public Safety	SCUT/HKU	2021
	GeoMAN [162]								General	XDU	2018
	Huang et al. [112]								Urban planning	PKU	2023
	Liang et al. [164]								Transportation	NUS	2021
	Balselbre et al. [13]								Urban Planning	NTU	2022
	Ruan et al. [235]								Transportation	NTU	2022
	Liu et al. [172]								Economy	HKUST(GZ)	2023
	PANDA [342]								Public Safety	XMU	2022
	UVLens [40]								Urban Planning	XMU	2021
	Miyazawa et al. [200]								Transportation	SUSTech	2019
	NodeSense2Vec [33]								Social	UCF	2021
	Keerthi Chandra et al. [134]								Urban Planning	UCF	2020
	Fu et al. [74]								General	UCF	2019
	Liu et al. [184]								Social	Gatech	2022
	Yuan et al. [349]								Transportation	RMIT	2021
	Bai et al. [12]								Transportation	Shanghai AI Lab	2019
	Ke et al. [132]								Transportation	Alibaba	2021
	Geng et al. [84]								Transportation	Alibaba	2019
	Yao et al. [329]								Transportation	PSU	2018
	Gao et al. [79]								Transportation	SWJTU	2022
	DeepMob [249]								Public Safety	SUSTech	2017
	Geng et al. [82]								Transportation	HKUST	2019
Alignment Based Data Fusion	Xi et al. [309]								General	THU	2022
	Zhang et al. [368]								General	THU	2021
	Yuan et al. [349]								Transportation	THU	2021
	Yin et al. [339]								Urban Planning	NUS	2020
	GSNet [272]								Public Safety	BJTU	2021
	Hashem et al. [98]								General	NTU	2023
	TrajGAT [328]								Transportation	NTU	2022
	RADAR [39]								General	XMU	2018
	Wang et al. [287]								General	CSU	2021
	Tedjopurnomo et al. [261]								Transportation	RMIT	2021
	ERNIE-Geol. [110]								General	Baidu	2022
	SAInf [193]								Transportation	JD Research	2023
	Gao et al. [77]								Transportation	SWJTU	2023
Contrast Based Data Fusion	KnowCL [186]								Economy	THU	2023
	Li et al. [154]								Economy	THU	2022
	MMGR [10]								General	NTU	2023
	ReMVC [366]								Urban Planning	NTU	2022
	HMTRL [173]								Transportation	UCF	2023
	Mao et al. [196]								Transportation	Shanghai AI Lab	2022
	UrbanSTC [226]								Transportation	JD Research	2022
	UrbanCLIP [323]								General	HKUST(GZ)	2023
Generation Based Data Fusion	SG-GAN [378]								Urban Planning	NUS	2020
	ActSTD [354]								Transportation	THU	2022
	DifFSTG [299]								General	BJTU	2023
	CP-Route [297]								Transportation	BJTU	2023
	G2PTL [304]								Transportation	Cainiao	2023
	DiffUFlow [401]								Transportation	CSU	2023
	DP-TFI [320]								Transportation	UESTC	2023
	Wang et al. [273]								Transportation	UCF	2021
	Chattraffic [358]								Transportation	BJUT	2023
	MGEO [61]								General	Alibaba	2023

ses such as traffic projections and socioeconomic pattern recognition. The distinctive feature of this category lies in directly merging features using methods such as addition, multiplication, concatenation, or graph-based operations. This type of fusion has relatively low computational complexity, primarily depending on the feature dimensions and the specific fusion operation, typically linear  $O(n)$ . Here,  $n$  represents the number of feature dimensions involved in the fusion process, indicating total number of features being combined from various sources.

**Alignment-based Data Fusion** aims to identify a shared feature space or structure among diverse data representations, enabling their alignment or integration. This involves the model learning to transform information from one source to another for semantic consistency. For instance, in tasks merging images and text, the model must align visual features with textual descriptions, often achieved through a multi-modal embedding space [229, 14]. The widely used attention mechanism [268] exemplifies alignment-based data fusion, assigning weights to different input parts to prioritize them. In tasks like image captioning, cross-modal attention aligns and weights specific image regions with the textual description [190, 361, 406]. The computational complexity could be higher due to extensive matrix operations, typically quadratic  $O(n^2)$ , where  $n$  is the input sequence length.

**Contrast-based Data Fusion** employs the contrastive learning framework to enhance feature discriminability at the sample level, contrasting with alignment-based data fusion that focuses on feature-level alignment [56, 202, 147]. By training the model to differentiate categories or samples, it identifies key distinguishing features. In urban computing, this involves comparing traffic patterns or environmental variables across different areas, time points, and modalities. The pairing of positive and negative samples enables the model to learn and excel in distinguishing complex urban situations, refining the acuity of computational tools [373, 258, 216]. Utilizing contrastive learning to enhance feature discriminability, this method requires significant computation for sample pairing and similarity calculations, with complexity often quadratic  $O(n^2)$ . Here,  $n$  represents the number of samples involved in the contrastive learning process, indicating the total number of data points being compared.

**Generation-based Data Fusion** utilizes deep learning's creative capacity to generate one urban modality under the condition of the same or other modalities [179, 71, 265]. In urban computing, this approach proves beneficial for simulating diverse scenarios, such as crafting traffic patterns under various circumstances and assessing urban planning outcomes [355, 5]. The generative methods involved in urban multi-modal fusion encompass mask modeling, diffusion, and LLM-enhanced techniques. Employing generative models such as GANs and VAEs to simulate urban scenarios, this approach involves highly complex training and optimization processes, leading to very high computational complexity, also generally quadratic  $O(n^2)$  or higher. Specifically,  $n$  indicates the number of data points or parameters involved in the training and optimization process of the generative model, indicating the scale of the data used to simulate urban scenarios.

## 4.2. Feature-based Data Fusion

Feature-based fusion integrates information from different modalities through methods like *feature addition*, emphasizing equal importance, and *feature multiplication*, highlighting joint significance (Sec.4.2.1). Additionally, techniques such as *feature concatenation* (Sec.4.2.2) and *graph-based fusion* (Sec.4.2.3) enhance multi-modal understanding by combining feature vectors or leveraging graph structures to represent inter-modal connections. This fusion approach is characterized by its simplicity and efficiency, making it particularly advantageous for applications requiring high real-time performance, such as transportation. However, it is relatively coarse and lacks the ability to capture and align details accurately, which may lead to underperformance in tasks necessitating the amalgamation of diverse datasets, such as urban planning.

### 4.2.1. Feature Addition and Multiplication

Element-wise addition is a fundamental operation among multi-modal fusion techniques, where corresponding elements of vectors or matrices from different modalities are summed to create a fused representation. Given two vectors  $\mathbf{X} = [x_1, x_2, \dots, x_n]$  and  $\mathbf{Y} = [y_1, y_2, \dots, y_n]$  representing features from different modalities, the element-wise addition operation results in a new vector  $\mathbf{Z} = [z_1, z_2, \dots, z_n]$ , where  $z_i = x_i + y_i$  for  $i = 1, 2, \dots, n$ . Mathematically, the element-wise addition operation is expressed as

$$\mathbf{Z} = \mathbf{X} + \mathbf{Y}. \quad (2)$$

Element-wise multiplication varies from the resulting new vector  $\mathbf{Z} = [x_1y_1, x_2y_2, \dots, x_ny_n]$ . Mathematically, such an operation can be expressed as

$$\mathbf{Z} = \mathbf{X} \odot \mathbf{Y}, \quad (3)$$

where  $\odot$  denotes element-wise multiplication. These fusion approaches provide a simple yet effective way to integrate information, preserving the individual contributions of each modality in the combined representation.

For example, Guo et al. [94] developed the ROD-Revenue framework to predict driver revenue using a linear regression model, given features extracted from multi-source urban data. The basic features from multiple datasets, including ride-on-demand (RoD) service, taxi service, and POI information, are used to construct composite features in a product form. Likewise, Bai et al. [12] fused the historical passenger demand, meteorological data, and time meta to learn a joint representation for citywide passenger demand prediction.

### 4.2.2. Feature Concatenation

Concatenation is a widely used technique for multi-modal fusion, involving the combination of feature vectors or matrices from different modalities by appending them along a specified axis. Let vectors  $\mathbf{X} = [x_1, x_2, \dots, x_m]$  and  $\mathbf{Y} = [y_1, y_2, \dots, y_n]$  represent features from two modalities. Concatenating these vectors along the concatenation axis results in a new vector

$\mathbf{Z} = [x_1, x_2, \dots, x_m, y_1, y_2, \dots, y_n]$ , forming a fused representation. Mathematically, the concatenation operation is denoted as

$$\mathbf{Z} = \text{concat}(\mathbf{X}, \mathbf{Y}). \quad (4)$$

This method allows the preservation of individual modality information while creating a comprehensive representation for downstream tasks.

For instance, ST-ResNet [362, 364] dynamically aggregated the outputs of three different residual temporal networks, assigning different weights to different branches and regions. The aggregation was further combined with external factors such as meteorological information (illustrated in Figure 9). Similarly, Lin et al. [170] proposed a DeepSTN+ framework to forecast urban crowd flows via long-range spatial dependence modeling and the introduction of prior knowledge such as POI distribution. The three categories of historical crowd flow maps - closeness, period and trend, are concatenated firstly, followed by convolution as the fusion of different kinds of information. Some studies [163, 213, 253] followed this concatenation scheme to fuse multimodal information (i.e., external factors) as well. Furthermore, Liang et al. [164] appended the same three temporal sequences with meta features representing POI and road network (illustrated in Figure 10), which serve as fine-grained knowledge for urban flow prediction.

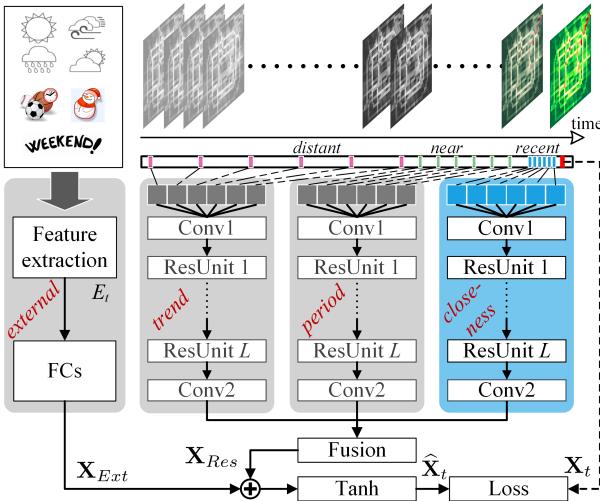


Figure 9: ST-ResNet dynamically fused the results from three distinct residual temporal networks, allocating varied weights to diverse branches and regions. This fusion process was additionally integrated with external factors, including meteorological information. [362, 364].

In addition, concatenation operation is often used in visual feature fusion. Huang et al. [113] concatenated the high-dimensional features of satellite imagery, street-level imagery, and taxi trajectory time series together, and then passed them through a softmax layer to distinguish between urban and non-urban villages. Yin et al. [337] also concatenated latent representations of satellite imagery and GPS trace as a fusion method for missing road attribute inference. In addition, the features of household capacity, human mobility, and commercial hotness can be merged together to estimate the population [40].

#### 4.2.3. Graph-based Data Fusion

Traditional deep learning-based feature extraction and fusion methods have brought revolutionary advancements to various urban computing tasks in recent years. These tasks typically involve data represented in the Euclidean space. However, there is a growing need for addressing tasks where data is generated from non-Euclidean domains and represented as graphs with intricate relationships between objects such as POI networks. In urban computing, graphs serve as representations of intricate networks that encompass various elements such as roads, buildings, and social interactions. In general, a graph can be represented as

$$\mathbf{G} = (\mathbf{V}, \mathbf{E}) \text{ (General form)}, \quad (5)$$

or

$$\mathbf{G}^{(t)} = (\mathbf{V}, \mathbf{E}, \mathbf{X}^{(t)}) \text{ (Spatio-temporal form)}, \quad (6)$$

where  $\mathbf{V}$  represents a set of nodes and  $\mathbf{E}$  is the set of edges. For a spatio-temporal graph, where the node attributes vary dynamically over time,  $\mathbf{G}^{(t)}$  is the graph representation at time step  $t$  and  $\mathbf{X}^{(t)}$  is the node feature matrix of graph  $\mathbf{G}$  at the same time step. Deep learning models based on spatio-temporal graphs have gained significant importance and achieved remarkable success across various applications [343, 158, 115, 308].

An intuitive approach for graph-based data fusion is **multi-view graph network-based data fusion**, which represents diverse data through a set of multi-view graphs [172, 74, 63, 82]. In this framework, each view is associated with different edge vectors, denoted as  $\mathbf{E}^i$ , which represent the relationships between each node (such POI), denoted as  $\mathbf{V}$ , under a particular feature view  $i$  (distance, mobility, semantic, etc.). For graphs from different views, graph nodes  $\mathbf{V}$  usually serve as connectors, linking them together for feature fusion. Each view in these frameworks captures a unique perspective or aspect of the underlying data, allowing for a comprehensive understanding of a more complex system.

Based on the conception of a multi-view graph, Fu et al. [74] proposed multi-view POI graph networks that incorporate various geo-features including regions, distances, and human mobility connectivity. Du et al. [63] developed a group of POI networks to characterize both static and dynamic features through the variable graph edges. Both of these work effectively integrate and fuse information from various data sources and achieve great success in downstream tasks. Liu et al. [172] conducted research on characterizing urban vibrancy by employing a multi-view graph framework and demonstrated the effectiveness of multi-view graphs in capturing the intricate relationship between urban vibrancy and urban spatiotemporal dynamics. In addition, different regions in a city can be represented through multi-view graphs as well. Geng et al. [82] proposed a multiple graph framework to encode pair-wise correlations between regions in urban areas. As shown in Figure 11, each region square, measuring  $(1km \times 1km)$ , is represented as a node in the graphs. By utilizing graph convolutional networks to observe and fuse these graphs, the framework can capture various region correlations and achieve impressive results in forecasting ride-hailing demand.

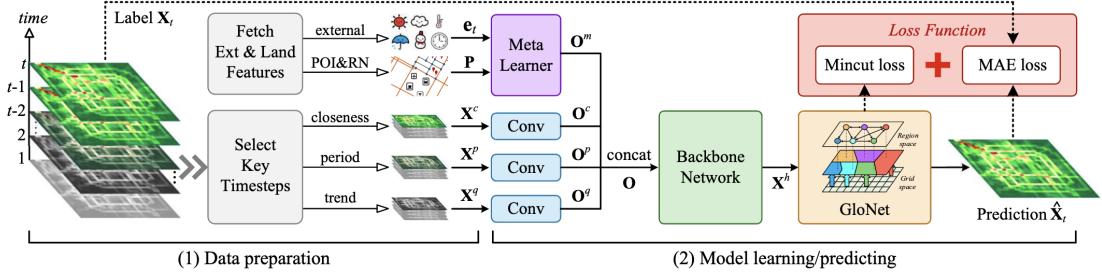


Figure 10: The feature-based data fusion framework proposed by Liang et al. [164]. Features extracted from the temporal sequence are concatenated together with external features from POI and road networks.

While it is efficient to incorporate features from different data sources using multi-view graphs, describing high-dimensional cross-modal correlation between different items can be challenging as there is no direct interaction between nodes from each graph. To address these difficulties, researchers proposed spatiotemporal heterogeneous networks (SHNs) to encapsulate the complex cross-modal relationships among different urban entities. Subsequently, **heterogeneous graph-based data fusion** is considered as a promising approach to the fusion of various data in one graph and tackles the above challenges [176, 365, 278, 134, 408, 292]. Liu et al. [184] proposed a hierarchical embedding framework that aimed to establish connections between nodes in the city activity graph and user interaction graph. Although the specific emphasis on SHNs may not be explicitly stressed, this attempt, along with the successful cross-modal node connection, has demonstrated significant success in capturing and representing cross-modal connections. Zhang et al. [365] proposed a compound multi-linear relationship graph network to converge various edge connectivity with interaction between multi-view graphs. In their methodology, which is shown in Figure 12, any random walk path in the compound graph is a compound of various kinds of relationships for different views. This work strongly develops the generality of graph neural networks in cross-domain data fusion.

Chandra et al. [33] designed a SHNs embedding framework that takes the POIs as nodes and human mobility between POIs as weighted links. The SHNs in their research are represented as

$$\mathbf{G} = (\mathbf{V}, \mathbf{E}, \phi, \psi), \quad (7)$$

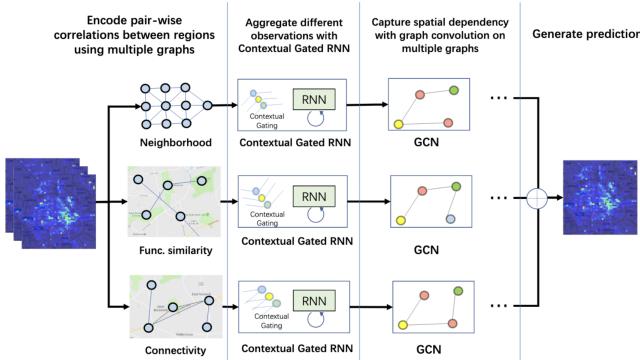


Figure 11: The overall architecture of the spatio-temporal multi-graph convolution network proposed by Geng et al. [82].

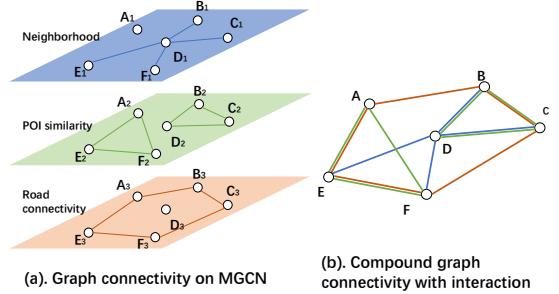


Figure 12: In (a), the multi-view graph connectivity is shown for each individual graph. Vertices represent regions, and the weighted edges represent region-wise relationships. There is no interaction between the graphs. In (b), the compound graph connectivity is represented as a multi-linear graph. Vertices are connected if there is an edge in any of the traditional multi-view graphs. This allows for a more comprehensive understanding of the interdependencies among regions across all graphs [365].

where  $\phi : \mathbf{V} \rightarrow \mathcal{L}$  is a mapping function for nodes and  $\psi : \mathbf{E} \rightarrow \mathcal{R}$  is a edge type mapping function. Within this presentation based on SHNs, the framework could analyze urban mobility data from multiple sources in a uniform model space.

In addition to the commonly researched SHNs, Liu et al. [176] proposed a multi-modal transportation graph consisting of nodes for users, transport modes, and origin-destination pairs. This approach aimed to develop a multi-modal planning methodology in a transport recommend system. Similarly, Wang et al. [278] constructed an urban knowledge graph that effectively integrates features and knowledge from various trajectory data sources to model users' mobility patterns.

#### 4.3. Alignment-based Data Fusion

Alignment fusion ensures that semantically related content across modalities is effectively combined. As illustrated in Figure 13, methods for multi-modal alignment span categories such as *cross-modal attention mechanism* (Sec.4.3.1) and *multi-modal encoder-based fusion* (Sec.4.3.2), enabling understanding and collaboration between disparate sources of information. Alignment-based approaches can achieve more precise modal alignment and exhibit flexibility, making them suitable for various general scenarios. However, they generally require higher computational resources, which makes them more appropriate for offline tasks in transportation and urban planning.

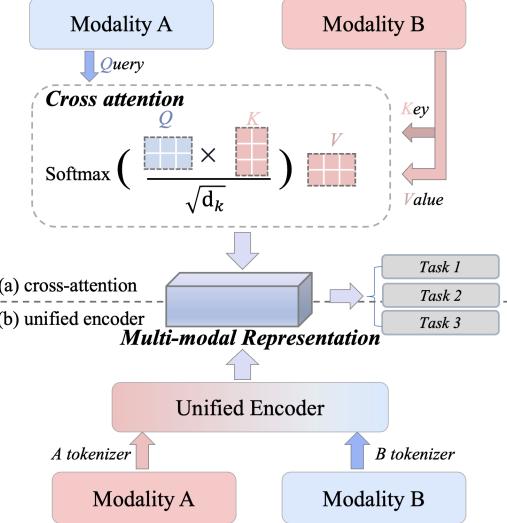


Figure 13: The general framework of alignment-based cross-domain data fusion in urban computing: (a) cross-attention framework for attention-based alignment; (b) unified encoder framework for encoder-based alignment.

#### 4.3.1. Attention-based Alignment

Attention mechanism [268], especially multi-modal cross-attention is a fusion technique crucial for integrating information across diverse modalities, such as text and images [296, 122, 327]. The cross-attention mechanism can be summarized in the following three steps. First, project the feature vectors of modalities  $\mathbf{X}$  and  $\mathbf{Y}$  into query ( $\mathbf{Q}$ ), key ( $\mathbf{K}$ ), and value ( $\mathbf{V}$ ) spaces using learnable parameter matrices  $\mathbf{W}_Q$ ,  $\mathbf{W}_K$ , and  $\mathbf{W}_V$ , respectively:

$$\mathbf{Q}_X = \mathbf{X}\mathbf{W}_{Q_X}, \quad \mathbf{K}_Y = \mathbf{Y}\mathbf{W}_{K_Y}, \quad \mathbf{V}_Y = \mathbf{Y}\mathbf{W}_{V_Y}. \quad (8)$$

Second, compute the initial attention scores by taking the dot product of the query vectors of modality  $\mathbf{X}$  (i.e.,  $\mathbf{Q}_X$ ) and the key vectors of modality  $\mathbf{Y}$  (i.e.,  $\mathbf{K}_Y$ ), divided by the square root of the dimensionality of the key vectors  $\sqrt{d_k}$ . Subsequently, we apply the softmax function to obtain normalized attention weights, ensuring that the weights sum up to 1:

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}_X \mathbf{K}_Y^T}{\sqrt{d_k}}\right). \quad (9)$$

Third, use the attention weights to compute the weighted sum of values for each modality:

$$\mathbf{Z}_X = \mathbf{A}\mathbf{V}_Y, \quad \mathbf{Z}_Y = \mathbf{A}^T\mathbf{V}_Y. \quad (10)$$

In recent years, the urban computing community leveraged such fusion mode and its variants for comprehensive urban modality alignment. For example, to model the relations among the target road attributes, Yin et al. [339] generated task-specific fused representations by applying attention-based feature fusion of location, bearing, speed, and map context. Zhang et al. [368] applied a GAT-based attention mechanism in learning region representations from two views of the built correlations (i.e., human mobility view and region attribute view), and a

joint learning module to fuse multi-view embeddings. In this proposed multi-view joint learning module shown in Figure 14, the self-attention layer enables information sharing across all views and the fusion layer is responsible for combining multi-view representations via adaptive weights.

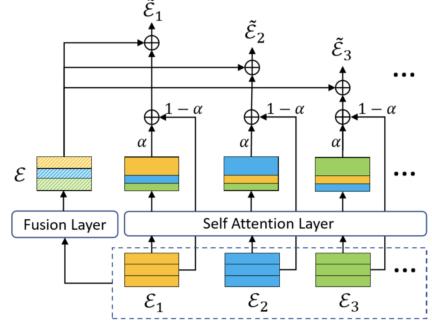


Figure 14: The architecture of multi-view joint learning module, consisting of a self-attention layer and a fusion layer[368].  $\mathcal{E}_i$  is the representation for the  $i$ -th view of global information, and  $\alpha$  is the weight of global information.

The MVMT-STN model [288] included a multi-view GCN to capture the global semantic dependency between POI, road network, and risk information. Huang et al. [110] used a transformer-based aggregation layer to model the graph structure containing both toponym and spatial knowledge. Qiang et al. [225] designed a transformer encoder to represent the features of each package by integrating those of other candidate packages. The RankETPA model [298] was designed for package pick-up arrival time prediction, and there is an attention module to model the interaction between the package's features and the courier's features.

The other similar attentional fusion work includes [309] (POI-view and geographic-view representations), [349] (region-level and inter-traffic correlations), [272] (geographical and semantic spatio-temporal representations) and [323] (satellite visual and textual representations).

#### 4.3.2. Encoder-based Alignment

Encoder-based fusion entails the integration of multiple modalities into a shared encoder architecture, as opposed to employing separate encoders for each modality. This approach leverages a unified deep learning model (e.g., RNN variants [248, 123] and self-attention methods [28, 156]) to collectively process and extract meaningful representations from diverse sources of information. By cohesively injecting multi-modal data into a single encoder, the model is inherently encouraged to capture intricate inter-modal relationships and dependencies.

For example, Song et al. [248] developed the DeepTransport model for mobility simulation and transportation mode prediction, where heterogeneous data including GPS records and transportation information are fed into a deep LSTM learning architecture. Such multi-layer LSTM has been demonstrated to be able to learn at different time scales over the input [103]. The DeepUrbanEvent system [123] contained a ConvLSTM [243] encoder module for simultaneous multi-step forecasting of crowd density and crowd flow. The ConvLSTM extends the

fully connected LSTM (FC-LSTM) to have convolutional structures in both input-to-state and state-to-state transitions. The RegionDCL framework [156] leveraged the Transformer encoder with average-pooling as the region-level encoder of building features and POI information. Cai et al. [28] proposed a multi-level graph encoder equipped with a GAT encoding module to capture couriers’ both high-level transfer modes between AOIs and low-level transfer models between locations.

#### 4.4. Contrast-based Data Fusion

Contrastive learning, a pivotal paradigm in machine learning, can be categorized based on contrast creation [147, 202]. The categorization encompasses methods such as instance contrast, batch contrast, and temporal contrast, which each focuses on contrasting representations through augmentations, batch comparisons, and temporal shifts, respectively. Compared to alignment-based methods, the contrast-based data fusion method enhances discrimination through negative sample augmentation. However, it imposes stringent requirements on the selection of negative samples and batch size. Due to its demand for large datasets, it is well-suited for downstream tasks where data is abundant or easily accessible, such as those related to transportation, environment and economy.

The InfoNCE (Noise-Contrastive Estimation) loss, a common objective function in contrastive learning, aims to maximize the similarity between positive pairs and minimize the similarity between negative pairs [210]. Mathematically, the InfoNCE loss is formulated as follows:

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{X}, \mathbf{Y}) = -\log \left( \frac{\exp(\text{sim}(\mathbf{X}, \mathbf{Y}))}{\exp(\text{sim}(\mathbf{X}, \mathbf{Y})) + \sum_{k=1}^K \exp(\text{sim}(\mathbf{X}, \mathbf{N}_k))} \right), \quad (11)$$

where  $K$  is the total number of negative samples,  $\text{sim}(\mathbf{X}, \mathbf{Y})$  is the similarity measure between positive pairs, and  $\mathbf{N}_k$  represents negative samples.

CLIP (Contrastive Language-Image Pre-training) is a typical model in contrastive learning, designed to concurrently learn representations of images and text [227]. CLIP can be considered a form of instance contrast because it encourages the model to align representations of a given instance across different modalities. The model achieves this by minimizing the negative logarithmic probability of similarity between positive image-text pairs while contrasting against negative pairs. Mathematically, the CLIP loss function ( $\mathcal{L}_{\text{CLIP}}$ ) is expressed as:

$$\mathcal{L}_{\text{CLIP}}(\mathbf{I}, \mathbf{T}) = -\log \left( \frac{\exp(\text{sim}(\mathbf{I}, \mathbf{T}))}{\exp(\text{sim}(\mathbf{I}, \mathbf{T})) + \sum_{k=1}^K \exp(\text{sim}(\mathbf{I}, \mathbf{N}_k))} \right). \quad (12)$$

Here,  $\mathbf{I}$  and  $\mathbf{T}$  represent image and text representations, respectively. In our survey, contrastive data fusion can be categorized into *instance contrast-based fusion* (Sec.4.4.1) and *batch contrast-based fusion* (Sec.4.4.2), as illustrated in Figure 15.

##### 4.4.1. Instance Contrast-based Fusion

In instance contrast, the model aims to fuse the knowledge between different views of the same instance. It is effective for capturing intricate details within individual data points, promoting the model’s ability to recognize fine-grained patterns and features [147, 56, 183].

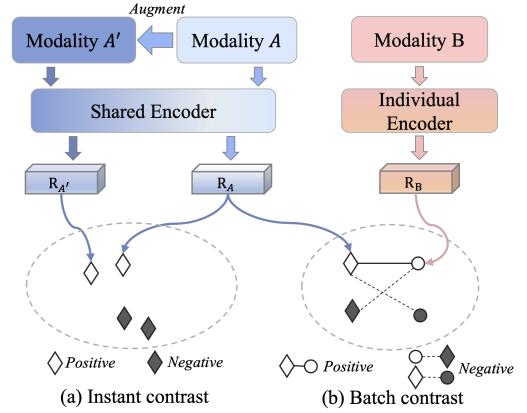


Figure 15: The general framework of contrast-based cross-domain data fusion in urban computing: (a) instant contrast; (b) batch contrast.

Inspired by the computer vision domain [319, 46, 407], urban multi-modal research started to construct self-augmented data as contrastive pairs. For example, Li et al. [154] explored self-similarity across urban images to construct contrastive samples via data augmentation methods including rotation, gray-scale, and flipping. Similarly, Bai et al. [10] learned the physical properties of the geographies via intra-modal contrast among very high resolution (VHR) image augmentations. In addition to imagery augmentation, Zhang et al. [366] designed three POI-level augmentations for intra-view contrastive fusion: random insertion, random deletion, and random replacement. Furthermore, Mao et al. [196] introduced domain-specific augmentations for road-road contrast and trajectory-trajectory contrast separately, i.e., road segment with its contextual neighbors and trajectory with its detour replaced and dropped alternatives.

##### 4.4.2. Batch Contrast-based Fusion

Batch contrast involves contrasting samples within the same batch. It introduces a form of global context, where the model learns to distinguish features not just within instances but also in relation to the entire batch [147, 56, 183].

Geographical similarity-guided contrastive learning is pivotal in urban computing because urban images adhere to Tobler’s First Law of Geography [198]. For instance, Li et al. [154] further enhanced the contrastive learning method by taking into account geographical similarity and minimizing the feature distance between two images that are geo-adjacent.

Besides, the CLIP-based paradigm has been explored in urban computing in recent years. For instance, Liu et al. [186] proposed the KnowCL model for socioeconomic prediction, which is the first solution that introduces the regional knowledge graph-based semantics and its associated imagery representation as contrastive pairs. Except for self-augmented contrast, Bai et al. [10] also bridged the socio-economic semantic gap through inter-modal contrast between VHR images and POIs. The RecMVC model of Zhang et al. [366] included an inter-view contrastive learning module between POI and mobility, serving as a soft co-regularizer to transfer knowledge across multi-views. Liu et al. [173] designed a trajectory con-

trastive learning paradigm, where hub and link representations in the same trajectory are enforced to have a higher correlation with one another. Likewise, Mao et al. [196] introduced road-trajectory cross-scale contrast to bridge the two scales by maximizing the total mutual information. This contrast is elaborately tailored via novel positive sampling and adaptive weighting strategies. Following the conventional CLIP setting, Yan et al. [323] leveraged satellite imagery and associated LLM-generated description as positive pairs, to learn a robust visual embedding for urban region profiling (depicted in Figure 16). Subsequent work, UrbanVLP [97], has broadened the contrastive learning paradigm to encompass multi-granularity visual clues, including satellite and street-view images. Concurrently, it also proposes effective methods to guarantee the generated text quality.

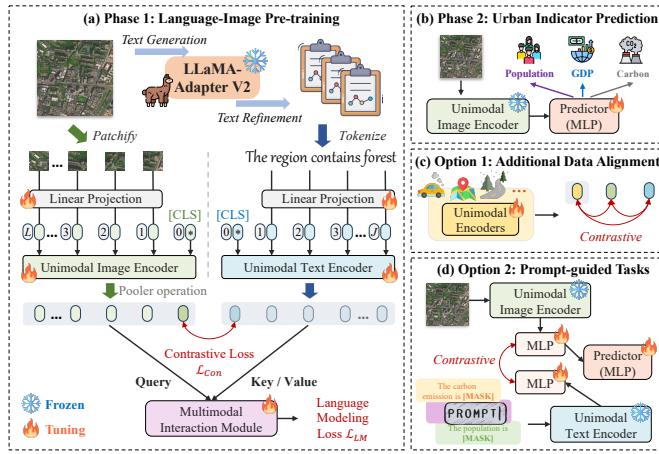


Figure 16: The framework of UrbanCLIP, the first-ever LLM-enhanced framework that integrates the knowledge of textual modality into urban imagery profiling. It utilizes the CLIP rationale for language-image pertaining [323].

#### 4.5. Generation-based Data Fusion

The primary objective of a generation-based model is to produce the desired data based on specified input conditions. Both input and output data may manifest in diverse formats [323, 356, 358], rendering it a favorable choice for data fusion. By generating new content that correlates with input data, generative models are driven to discern intricate correspondences between multimodal information, thereby facilitating efficient information aggregation. However, these models are marked by significant computational complexity and considerable training challenges. Consequently, they are not suitable for tasks requiring high real-time performance. Instead, they excel with large datasets [131], making them particularly well-suited for industries such as transportation, economics, and energy. Compared to the general Computer Vision (CV) or Natural Language Processing (NLP) area, the field of urban computing exhibits a relatively limited volume of generative research contributions. However, in recent years, there has been a growing number of generative works emerging, particularly influenced by the rise of LLMs [180]. Based on the specific methods of generation, here we categorize the generation-based data fusion into four types, including autoregressive (Sec.4.5.1), mask

modeling (Sec.4.5.2), diffusion-based (Sec.4.5.3), and LLM-enhanced models (Sec.4.5.4).

##### 4.5.1. Autoregressive Model

The autoregressive model was first developed for the Language Modeling (LM) tasks in NLP, which predict future data based on historical data. Given a text sequence,  $x_{1:T} = [x_1, x_2, \dots, x_T]$ , the learning objective of a language model is to maximize the probability of a sequence, which can be mathematically formulated as:

$$\mathcal{L}_{AR} = \max_{\theta} \sum \log P_{\theta}(x_t | x_{t-k}, \dots, x_{t-1}), \quad (13)$$

where  $k$  is the size of the sliding window.

Gradually, the definition of input data and output data for autoregressive model has expanded from text to encompass multi-modal data. Owing to the multi-modal nature of the attention mechanism [268], whereby queries, keys, and values can stem from different modalities, it has facilitated the exploration of multi-modal fusion. VirTex [57] proposed that compared with contrastive learning which uses classification labels as a learning signal, captions can provide a more semantically dense learning signal. SimVLM [295] reformulated the typical encoder-decoder architecture for end-to-end training with a single prefix language modeling objective, achieving the efficient utilization of weakly aligned image-text pairs. CoCa [344] is a follow-up work to ALBEF [152] and SimVLM [295], which creatively integrated attentional pooling methods in a creative manner, combining contrastive and generative approaches to achieve exceptional performance.

In the field of urban computing [391], the generative decoder has played a significant role, even before the era of transformer models [268]. GeoMAN [162] first introduced a multi-layer attention mechanism for spatio-temporal data prediction based on the encoder-decoder architecture. The generative decoder combined LSTM and Temporal Attention to predict the future performance of sensors. It achieved excellent performance in applications such as water quality prediction and air quality. Zhao et al. [382] implemented a bottom-up and top-down framework to do street-view image classification, which utilized RNN units to make predictions based on visual elements and contextual information. Besides, in urban computing area [391], the limited availability and diverse nature of domain data, as well as the challenges in its acquisition, have led to increased interest in multi-modal pretraining in recent years.

##### 4.5.2. Mask Modeling-based Fusion

The concept of masking modeling task has its origins in the field of NLP, contributing to the success of BERT [60]. MAE [102] is the first work that introduces masking modeling into CV. Compared with masking modeling works in NLP, MAE [102] has a larger mask ratio due to the fact that the images have less information density than natural language [102]. Graph-MAE [104] unleashes the power of mask modeling for graph structure. Different from most graph autoencoders' efforts in structure reconstruction, it proposes to focus on feature reconstruction with both a masking strategy and scaled cosine error. In the context of mask modeling structure for multi-modal

data fusion, various modalities are typically provided as input, followed by the masking of a portion or all of the modal data at varying proportions. Subsequently, information from other modalities is integrated to reconstruct the masked data [142, 83, 61].

MGeo [61] combined text and geolocation to implement location embedding for query-POI matching, which utilized two unimodal masked language modeling (MLM) loss and one Multi-Modal MLM loss to enforce information interaction. As illustrated in Figure 17, G2PTL [304] is a graph-based pre-trained model for text-based delivery address embedding. It leveraged MLM as one of the pre-training tasks to simulate input noise in missing or incorrect address cases in real-world scenarios. QUERT [314] adapted pre-trained language models to specific Travel Domain Search applications, which designed a Geography-aware masking strategy to force the pre-trained model to pay more attention to the geographical location phrases. ERNIE-GeoL [110] was dedicated to integrating toponym knowledge and spatial knowledge. By utilizing a Masked Language Modeling training procedure, it can acquire four types of toponym knowledge, including relationships between POI name, address, and type.

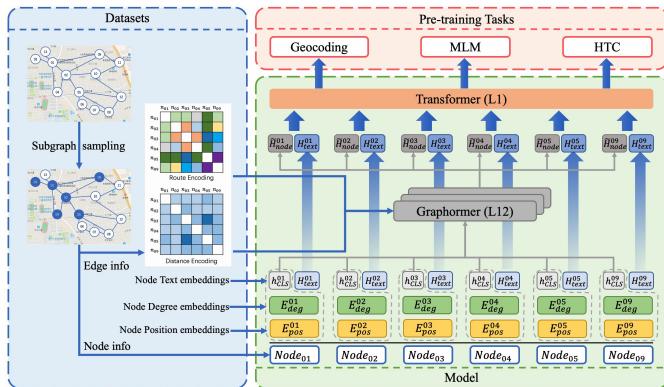


Figure 17: The framework of G2PTL, a Geography-Graph Pre-Trained model for delivery address in the Logistics field, which uses the MLM task to learn the semantic information in the address [304].

In the remote sensing area, despite the similarities between remote sensing images and RGB images, they also possess distinctions, with remote sensing images containing geographical information and temporal tags, as well as spectral multiple channels, and diverse resolutions and scales. SatMAE [49] discussed the domain-specific temporal and spectral characteristics of satellite imagery, and device’s temporal encoding and spectral positional encoding respectively to adapt MAE structure to satellite imagery representation learning. Scale-MAE [230] was designed to address another challenge in remote sensing satellite imagery: the variation in image scale due to data from multi-scale sensors. The pretraining process aims to learn a better representation of multiscale tasks by reconstructing low and high frequency features at different scales.

#### 4.5.3. Diffusion-based Fusion

In recent years, diffusion models [325], as an emerging and potent type of deep generative models, have demonstrated state-

of-the-art performance across various modalities such as images, speech, and video [30]. Diffusion models can serve as an appropriate framework for data fusion, enabling the seamless incorporation of new modalities due to the existence of conditions [401, 136]. Simultaneously, the integration of multiple modalities can also enhance the quality of generation.

Urban imagery, as the primary provider of visual information, has borrowed numerous techniques from the general computer vision area and achieved successful applications. DiffusionSat [136], inspired by Stable Diffusion [234] and ControlNet [367], provided the first large-scale generative foundation model for satellite imagery. It combined data from different modalities including geospatial metadata (latitude, longitude, ground-sampling distance), timestamp, and texts, achieving promising performance on a series of tasks such as super-resolution, temporal generation, and in-painting. The generation of street-view images [80, 276] for spatio-temporal applications, however, remains an underexplored area, which represents a promising research direction in the future.

Aside from common modalities like image, text, and audio, the spatio-temporal modality data covers a broader range of data formats in urban computing, including traffic situations [358], urban flow [401, 320, 299], and trajectories [356, 354]. In this context, the flexibility of diffusion models allows them to potentially exert a more considerable influence.

ChatTraffic [358] creatively introduced fine-grained text in the Text-to-Traffic Generation (TTG) task to adapt to unusual events. Besides, when enhanced with GCN to incorporate the spatial information inherent in the road network, it achieved more accurate and realistic long-term prediction.

DiffSTG [299] is the first work that generalizes the diffusion model DDPM to spatio-temporal graphs. As illustrated in Figure 18, it combined the spatio-temporal learning capabilities of STGNNs with the uncertainty measurements of diffusion models. The collection of fine-grained urban flow data is widely recognized as challenging due to the high costs associated with deployment and maintenance, as well as the presence of noise. Diffusion models are suitable tools to generate fine-grained flow maps from the coarse-grained ones [401, 320]. DiffUFlow [401] is the first generative approach for fine-grained urban flow inference. An ELFetcher module is proposed to utilize various external factors and land features as conditional guidance for the reverse denoising process.

Trajectory data represents another essential type of spatio-temporal data. However, it often raises privacy concerns due to the inclusion of personal geolocation information. One promising solution for this challenge is trajectory generation [356, 354], which aims to generate high-fidelity, privacy-free trajectories. The diffusion model, as a more reliable and robust method of generation than canonical methods, begins to be increasingly explored. DiffTraj [356] is the first exploration of trajectory generation by the diffusion model, which proposed a Traj-UNet architecture to predict the noise of each diffusion time step. It proved that the step-by-step denoising process of the diffusion model is an appropriate choice due to the stochastic and uncertain characteristics of human activities.

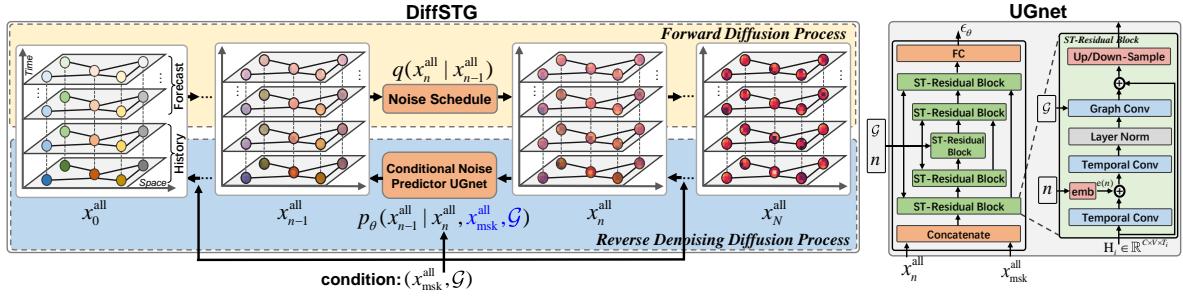


Figure 18: Illustration of DiffSTG and denoising network UGnet, which leverages an Unet-based architecture to capture multi-scale temporal dependencies and the Graph Neural Network (GNN) to model spatial correlations [299].

#### 4.5.4. LLM-enhanced Data Fusion

With the rise of GPTs [23, 214, 24], the remarkable capabilities of LLMs across a broad spectrum of fields have garnered significant attention, sparking a wave of research paradigm shift and shedding light on artificial general intelligence (AGI). The academia usually terms “large language models (LLM)” for these large-sized PLMs [388] due to the scaling laws [131] of Transformer [268] architecture. LLM-enhanced data fusion can be considered as a specific instance of encoder-based alignment, employing LLMs for feature encoding and information interaction. The extensive parameter size of LLMs endows it with potent alignment capabilities. LLMs have multifaceted impacts on the field of urban computing, including works that utilize LLMs’ geospatial capabilities [232, 195, 291, 322], applications in remote sensing [105, 141], works in the time series domain [34, 126, 29, 181, 402], etc, as illustrated in Figure 19.

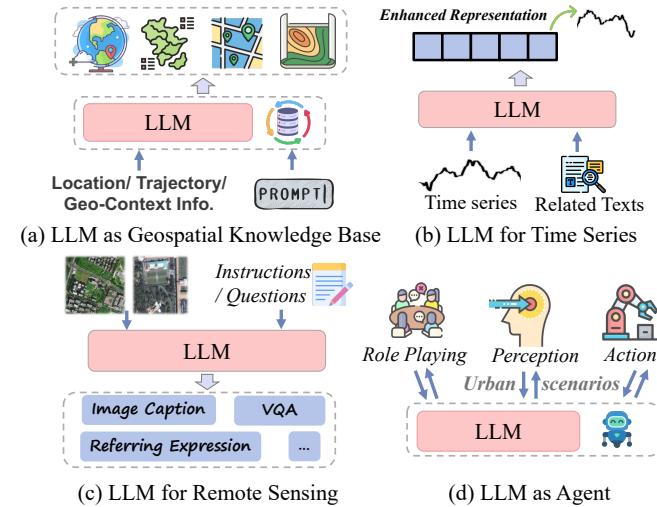


Figure 19: Categories of LLM-enhanced data fusion.

LLMs are able to compress and store geospatial knowledge within the training data. Exploring how to effectively utilize this compressed knowledge and design prompts that stimulate it are areas of current research interest [232, 195, 291, 322]. GPT4GEO [232] provided a comprehensive investigation of the extent to which GPT-4 [24] has mastered factual geographic knowledge and its ability to use this knowledge for reasoning.

GeoLLM [195] proposed that constructing the right prompt is key to extracting geospatial knowledge. Through providing geo-context information near a specific location, LLMs can be fine-tuned to achieve state-of-the-art performance on a variety of large-scale geospatial datasets for tasks such as predicting population density, house price, women’s education, etc. LLM-Mob [291] demonstrated the first attempt to apply LLM to modeling human mobility, which reformulated mobility data through historical stays and context stays to introduce long-term and short-term dependencies for prediction and reasoning.

Large Multi-modal Models (LMMs) [53, 174, 332, 369] demonstrate substantial efficacy in information fusion and modal alignment. In the domain of urban computing, the utilization of remote sensing imagery is prevalent for integrating visual elements from a bird’s-eye view perspective. Inspired by InstructBLIP [53], RSGPT [105] utilized frozen Image Encoder and LLM, while training a lightweight Q-Former [151] to align the two, achieving state-of-the-art remote sensing image captioning and remote sensing visual question answering downstream tasks. Developed based on LLaVA-v1.5 [174], GeoChat [141] introduced multi-modal instruction-tuning into the remote sensing domain and proposed the first versatile remote sensing Large Vision-Language Model with multitask conversational capabilities.

LLMs have the potential to revolutionize time series analysis [127, 165]. Following the success of large foundation models in NLP and CV, there are desires in the time series forecasting area to utilize pre-trained LLMs as powerful representation learners [34, 126, 29, 181, 402]. LLM4TS [34] combined patching and channel-independence techniques with temporal encoding, which unlocked the flexibility of pre-trained LLMs without introducing large parameter overhead. As depicted in Figure 20, Time-LLM [126] solved the alignment of time series data and natural language modality to unleash the power of LLMs for time series forecasting through Prompt-as-Prefix or Patch-as-Prefix, which reprogrammed time series into text prototype representations.

LLMs approach real-world cross-domain data fusion in the form of agents [127, 310]. In human perception of the world, diverse modalities ultimately converge into language, serving as the medium for expression and communication [239]. Based on essential language capabilities, LLMs can store knowledge and process cross-modal information, thereby continuously re-

ceiving feedback and interacting with the environment, demonstrating potential in spatiotemporal domains for comprehensive multimodal data integration. Zhou et al. [404] first explores urban planning through multi-agent collaboration framework, partially demonstrated performance surpassing that of human experts. LLMLight [144] transcends the previous paradigm of using LLMs as assistants to enhance decision-making, by directly employing LLMs as Traffic Signal Control (TSC) agents for decision-making.

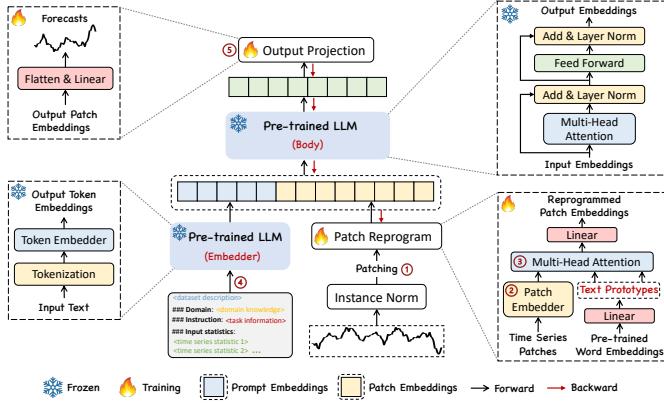


Figure 20: Time-LLM begins by reprogramming the input time series with text prototypes before feeding it into frozen LLM to align the two modalities [126].

## 5. Application Perspective

In this section, we categorize and summarize related works from the application perspective. Figure 22 illustrates applications from seven domains, including urban planning, transportation, economy, public safety and security, society, environment, and energy. Subsequently, we provide a comprehensive exposition of these applications, delving into intricate details, while also highlighting the pivotal role that deep learning-based data fusion methods have played in facilitating these tasks.

### 5.1. Urban Planning

Sensing a city and making effective planning and governing is of great importance to its development. Every political planning and strategy necessarily needs to be formulated under strong computational support which covers lots of factors, such as road network structures, geographical limitations, transportation situation, human mobility, and society. In earlier years, to formulate a decent planning or strategy for a city, planners always need to sensor the city through various labor-intensive surveys. Besides, the processing of these surveys from different sources was also tough to conduct.

With the development of city infrastructures, most urban data and factors can be directly collected from various sensors and platforms. However, processing and understanding such big data from multiple sources becomes a challenge due to the poor performance of traditional data analysis methods on big data. Deep learning models provide researchers with opportunities to handle big and multi-source data with better efficiency

and deeper understanding. Therefore, for most urban planning tasks leveraging the advantages of multi-source data, deep learning-based data fusion methods have become a strong support from urban planners in both city-level planning and region-level planning (shown in Figure 21).

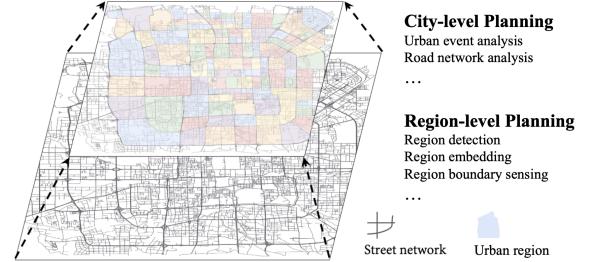


Figure 21: A city exhibits a multi-level structure for urban planning: city-level planning and region-level planning [154].

#### 5.1.1. City-level Planning

Multi-sources urban data provides comprehensive information about the operation and variation of cities. Deep learning-based data fusion could utilize the data to sense the city-level variation and understand it as well. This enables urban planners and researchers to understand urban dynamics, such as sensing social events [123, 384, 92], region's prosperity, city's vibrancy [177] and any important change happened to the city [13]. To sense city-level dynamics, Liu et al. [177] conducted research for understanding and predicting urban vibrancy evolution based on three data sources: mobile check-in data, POI data, and geographical data. Balsebre et al. [13] proposed a framework named *Geo-ER* to match geo-spatial entities from various data sources. [397, 395] introduced deep reinforcement learning to autonomously generate road layouts, aiming to connect various locations at the lowest construction expense.

**Urban event analysis** is crucial for urban planning as variations around a city usually come with urban events. Zhao et al. [385] proposed a multi-task learning framework to sense and predict urban events from the variation on social media. This framework could effectively train forecasting models for multiple locations simultaneously with shared information by restricting all positions to select a common set of features. Jiang et al. [123] proposed a recurrent neural network (RNN) based online system named *DeepUrbanEvent* to understand the crowd dynamic variation for social events. By extracting the deep trend from the current momentary observations on the historical GPS data from citizens, this system could generate an effective prediction for the crowd dynamic trend during a short future time at a big event in a city.

Focusing on the advantages of multi-source data, Zhao et al. [384] summarized the previous research in event prediction and stressed four challenges for variation understanding and event prediction from multi-source data: 1) geographical hierarchies; 2) hierarchical missing; 3) feature sparsity; 4) difficulty in update with incomplete multiple data. Subsequently, they proposed a multi-source feature learning model based on an Nth-

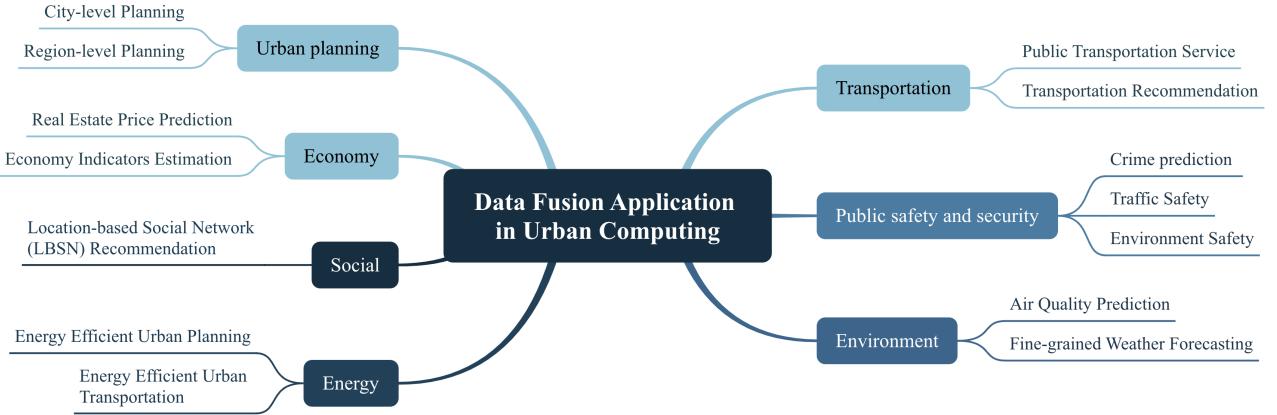


Figure 22: Taxonomy of application (category) and common downstream tasks (sub-category) for cross-domain data fusion in urban computing.

order geo-hierarchy and fused-overlapping group Lasso to handle these challenges. In their methodology, models can be instantly updated from new data from every data source with affordable computational cost, so the urban planner can respond to variation faster with this system.

**Road network analysis** plays a crucial role in the realm of urban planning, demanding significant attention from professionals in the field. The development of cities leads to continuous variation in the road conditions around a city. It is noteworthy that traditional road maps are unable to cope with these variations. Additionally, conventional surveying methods and database management techniques are inadequate for capturing and managing the rapidly evolving road networks. Yin et al. [339] pioneered the utilization of multi-source information for the automatic derivation of road attributes. They decided on a multi-task learning framework to extract low-level feature embedding from every data source and applied attention-based fusion to fuse the representations. Based on this work, they could combine the information from GPS trajectory and existing map data and achieve a significant improvement in classification on the OpenStreetMap data in Singapore. Based on the conception of data fusion, Yang et al. [324] proposed a *DuARE* system for large-scale automatic extraction by leveraging the GPS trajectory and the satellite images. Driven by its promising performance, DuARE has been deployed in China and has been updating the national road network by 100,000 km every month.

### 5.1.2. Region-level Planning

Cities serve as the foundation for societal functions in modern society. Based on the planning efforts of urban managers or the natural social operating principles, cities always develop into distinct functional regions, such as industrial zones, residential areas, and financial districts. These regions are sometimes referred to as Areas of Interest (AOIs). Simultaneously, within these areas, there are diverse locations that cater to the specific needs and demands of residents, commonly known as Points of Interest (POIs). These functional regions and points fulfill the daily requirements of citizens, enabling efficient urban operations. Accurately perceiving the status of AOIs and

POIs in a city and understanding them is crucial for urban planners and administrators in making informed decisions. This encompasses activities such as the selection of appropriate park locations, the planning of transportation routes, and the development of regional policies.

**Region detection** is the first step for us to understand region-level information. For POI or AOI detection, Xiao et al. [313] proposed a *Contextual Master-Slave Framework (CMSF)* that unitizing graph neural network to fusion the POI information and satellite image to detect urban villages. Huang et al. [112] fused satellite image, and street-level image with mobility data in their study, and the extra vision data were embedded through a *Vision-LSTM* network and were demonstrated crucial for region detection through comprehensive ablation study. Their work achieved an overall accuracy of 91.6% in identifying urban villages in Shenzhen, China. Zhao et al. [381] carried out a research on integrating information from social media such as Twitter as well as real-world locations. They designed a supervised Bayesian Model (sBM) to analyze the textual information, spatial features, POI information, and user behaviors. Based on this work, researchers are able to find user interests in special regions and understand the regions' properties.

**Region embedding** based on multi-source data is the foundation of region-level urban computing. Fu et al. [74] proposed a multi-view POI network by varying the representation of edges to fusion the geographical distance information and human modality. There are 2 kinds of POI-POI graph networks (shown in Figure 23) in this specialized multi-view framework: 1) distance-based POI-POI networks: *the weight of an edge represents the distance between two associated POIs*; 2) mobility connectivity-based POI-POI networks: *the weight of an edge represents the mobility connectivity between two associated POIs*. Zhang et al. [368] fused human mobility information with region attributes information with a multi-view joint learning module to infer the land usage of a city and Zhang et al. [366] designed a multi-task learning framework called *ReMVC* to classify regions by land usage and Li et al. [156] conducted a research on fusion publicly available building footprint information on *OpenStreetMap* and POI data for region representa-

tion learning. They carried out an experiment in Singapore and New York and demonstrated its efficiency.

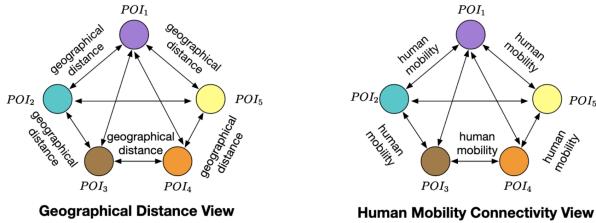


Figure 23: An example of the multi-view POI graph networks proposed by Fu et al. [74].

Bai et al. [10] also carried out research on geographic representation learning. In their methodology, satellite images and POI information are fused to solve multiple geographic mapping tasks. Besides, Bing et al. [20] presented a POI embedding method named CatEM, which jointly considered the spatial information and mobility information of POIs and achieved an impressive performance on POI classification tasks. Through efficient region embedding, Wang et al. [273, 275] proposed a generative framework to generate land-use configuration of a targeted region for urban planning.

**Region boundary sensing** plays a crucial role in region-level urban planning as it encompasses not only the political boundaries of the city but also functional or cultural regions within the city. Traditional administrative boundaries, often determined by policies and planning, may not accurately reflect the actual natural boundaries of a city.

Only a few deep learning researches were proposed for region boundary sensing in recent years [294, 37, 40, 270]. The natural boundaries of a city are constantly evolving and adapt to the real needs and functions of the urban area. In fact, accurately sensing these boundaries is crucial for formulating regional policies and adjusting administrative planning to align with the true dynamics of the city. Wang et al. [294] conducted a study on mapping the urban boundary of Zhengzhou City, China with its satellite images and POI data and found the result is in great agreement with the boundary ground truth.

Chen et al. [37] designed a cross-city federated transfer learning framework named *CcFTL* that can deal with multi-source data including POIs, road networks, and population density. In their methodology, this framework could not only deal with information from a single city but also transfer within various cities with great robustness. Besides, Chen et al. [40] proposed a boundary sensing framework for urban villages that fuse the information from satellite images, mobility from bike-sharing drop-off data, and POIs. This platform was successfully deployed on the government data platform of Xiamen City, China to serve both urban planners and citizens. Some researchers focused on sensing the boundaries of POIs, Vu et al. [270] proposed a boundary estimation frameworks by pairing the geo-tagged text information on tweets with the real name of POI and analyzing their spatial correlations. This work demonstrated that the spatial distribution of relevant tweets on the platform could reflect the social boundary of the targeted POI.

## 5.2. Transportation

Transportation serves as the arteries of a city, acting as bridges that connect different urban entities [62]. As a result, the comprehensive utilization of multi-modal data focuses a significant proportion of attention on downstream transportation-related tasks [199, 27, 245, 148].

### 5.2.1. Public Transportation Service

The advancement of a city is primarily demonstrated by enhancements in its transportation systems [62, 91, 52]. Public transportation, in comparison to private transportation, is recognized for its superior environmental sustainability and energy efficiency [199]. Guided by government supervision, public transportation embodies the features of public goods. In addition to fundamental passenger transportation, there are also freight transportation services, along with emerging services such as food delivery and ride-hailing, all of which contribute to enhancing urban convenience. Different from [391], where transportation systems were categorized based on the type of vehicles used, this classification of services is based on their specific application scenarios. Public transportation applications can primarily be categorized into three main groups: Safety, Flow Control, and Efficiency.

**Safety** is the most fundamental and urgent requirement. Consequently, roads, being high-risk areas for public transportation safety, have prompted numerous research efforts focusing on road safety. PANDA [342] predicted road risks after natural disasters by combining trajectories data and road event data. To identify road obstacles such as fallen trees and ponding water, RADAR [39] utilized co-training and active learning to fuse the heterogeneous features from trajectory data and environmental sensing data. Yin et al. [337] proposed a multi-modal fusion network for inferring missing road attributes. In their methodology (shown in Figure 24), the robustness of the network is largely enhanced by the pixel-level fusion of GPS traces and satellite images.

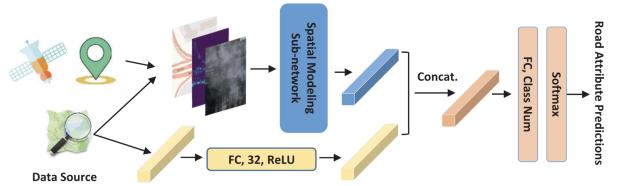


Figure 24: Network architecture of the proposed multi-modal fusion network by Yin et al. [337] for robust road attribute detection.

With the rapid urbanization of numerous countries, abnormal traffic incidents, have emerged as a substantial threat to public health and development. [287] used a Multi-View Multi-Task Spatio-Temporal Network to capture the various context features such as weather, POIs, and road conditions, which result in the occurrence of traffic accidents. The multi-task learning framework is effective in modeling features of different granularity levels, thanks to the tie of spatial associations.

Human factors play a critical role in transportation safety and can be considered a primary factor in accidents [205, 201].

SAX-DF [178] is a Symbolic Aggregate approximation Data Fusion model to comprehensively utilize multi-source and heterogeneous data, which can effectively improve the driver behavior detection performance with the help of a Positive Danger Mapping algorithm. The rise of autonomous driving has stimulated the demand for real-time driver behavior detection. Huang et al. [111] proposed a deformable inverted residual network to adaptively detect real-time driver behavior.

**Flow control** is an important downstream application that holds great significance for public policies such as resource scheduling and urban planning [86]. For instance, utilizing urban flow data enables governments to implement flow control measures during events such as New Year’s Eve at Shanghai’s Bund or Times Square in New York. By managing the crowd flow and directing subway passengers to nearby stations, the government can prevent potential dangers and promote efficient social functioning, thus yielding social benefits. Private ride-hailing platforms like DiDi and Uber allocate more transportation resources to densely populated areas, which allows them to gain better economic benefits.

Indeed, flow control can be broadly categorized into two main areas: i) Traffic Flow Control [17], which involves regulating and optimizing the flow of vehicles on roads and highways. ii) Mobility Prediction [360, 15], which endeavors to forecast and comprehend patterns of mobility behavior exhibited by humans or vehicles, is intended to enhance the planning and management of urban transportation and resources.

DeepSTN+ [170] utilized a spatio-temporal network to predict inflow and outflow regarding POIs and historical data. To infer urban flow, UrbanSTC [226] implemented contrastive pre-training in both spatial and temporal to learn robust features in both two modalities with different pretasks. Likewise, CSST [133] made use of a pretrain-finetuning paradigm to align low-quality GPS reports and external factors with the crowd flow.

For mobility prediction, [200] combined GPS traces and geo-tagged tweets to model human crowd flow. GraphTUL [79] solved the Trajectory user linking (TUL) problem by training an adversarial network in a semi-supervised way.

**Efficiency** primarily involves the promotion of efficiency in the allocation of transportation resources. It mainly relates to taxi/ride-hailing demand prediction, delivery time prediction, and passenger demand prediction.

DMVST-Net [329], shown in Figure 25, combined multi-view information to implement taxi demand prediction. It splits the process into spatial, temporal, and semantic views, which are modeled by local CNN, LSTM, and semantic graph embedding, respectively. To address the challenge of predicting travel demands in city regions for future time intervals demands, DeepTP [349] proposed to encode and capture three key properties from traffic data: Region-Level Correlations, Temporal Periodicity, and Inter-Traffic Correlations. To jointly predict demands for various ride-hailing service modes like solo and shared rides, Ke et al. [132] combined multi-graph convolutional networks with two multi-task learning structures, enhancing prediction accuracy for diverse ride-hailing services in urban areas. For delivery tasks, MetaSTP [235] introduced a meta-learning-based neural network model for predicting ser-

vice time in last-mile delivery. Utilizing a Transformer-based layer and location prior knowledge, MetaSTP addresses complex delivery scenarios, showing significant improvements in prediction accuracy and practical deployment in JD Logistics.

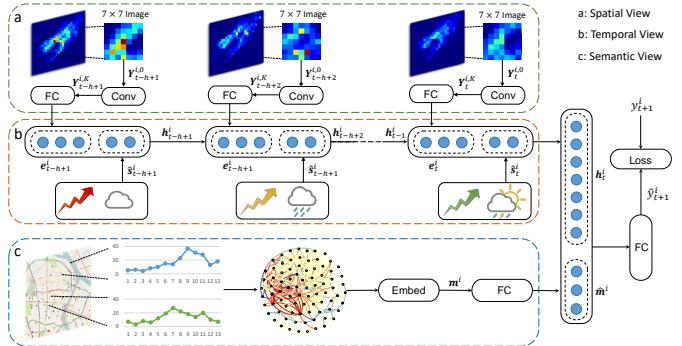


Figure 25: The architecture of DMVST-Net [329], which is designed to incorporate spatial, temporal, and semantic components for taxi demand prediction.

The prediction of passenger demand is also of great significance, which can assist transportation departments and related enterprises in enhancing the planning and management of passenger traffic. Bai et al. [12] proposed a deep learning framework combining graph convolutional recurrent neural networks and LSTM networks. This framework aims to accurately predict citywide passenger demand in ride-sharing platforms by analyzing historical demand data and external factors like weather and time. STG2Seq [11] solved the challenges for multi-step passenger demand forecasting in cities. It combines a Graph Convolutional Network (GCN) with an innovative encoder and attention-based output module, effectively capturing spatio-temporal correlations in-demand data and outperforming traditional methods in real-world tests.

### 5.2.2. Transportation Recommendation

Different from public services, **transportation recommendations** are proposed for specific personalized needs, significantly facilitating human society’s daily life, thus receiving increasing attention in recent years. CondorFerries [48] demonstrated that the prevalence of private travel has been on the rise since 2016, with 83 million Americans looking to take a solo trip in 2023. With the progress of human society and the development of urbanization, transportation recommendations are expected to create more economic and social benefits in the future. Among the various applications of transportation recommendations, route Recommendations, multi-modal transportation recommendations, and trip recommendations have garnered the most research attention. In this section, we will primarily introduce them respectively.

**Route recommendation** plays a core role in many applications such as taxi services like RoD (ride-on-demand) and navigation [157, 52]. Guo et al. [93] presented a novel force-directed algorithm for improving route recommendations in ride-on-demand services. This method, inspired by electrostatic principles, utilizes urban data to align vehicle distribution with passenger demand, thereby enhancing route efficiency.

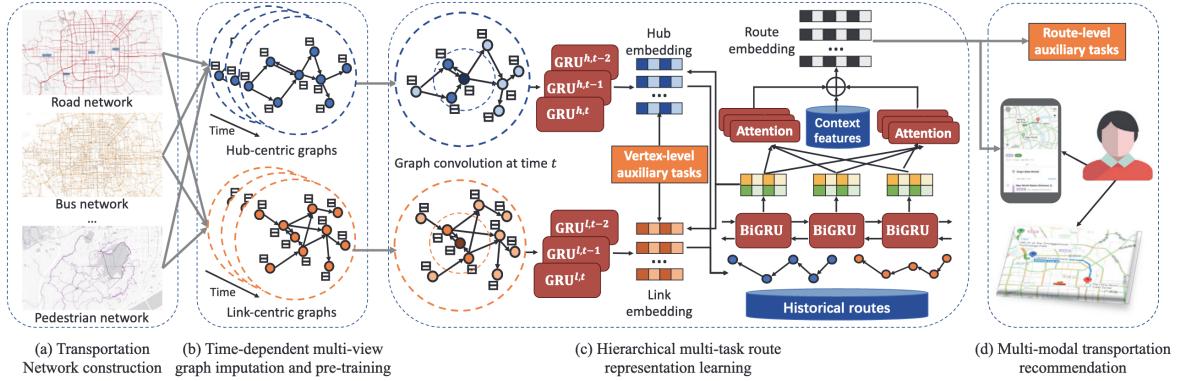


Figure 26: An overview of unified route representation learning for multi-modal transportation recommendation [173].

Given the origin and destination, **multi-modal transportation recommendations** offer travel plans consisting of diverse transportation modes (e.g., driving, cycling, and public transit), as well as instructions for making seamless transitions between different modes. Doing so improves individuals’ convenience and well-being, particularly for older adults and children. Liu et al. [176] proposed Trans2Vec, a model that enhances transportation mode recommendations by integrating heterogeneous data to capture user and location preferences, thereby improving accuracy in suggesting multi-modal transportation options. Liu et al. [173] developed a novel framework (shown in Figure 26 for multi-modal transportation recommendations. The approach leverages a hierarchical multi-task route representation learning technique that integrates spatio-temporal autocorrelation modeling and coherent-aware attentive route learning. Enhanced by spatio-temporal pre-training strategies, the framework effectively utilizes time-dependent multi-view transportation graphs, offering a more nuanced and accurate transportation recommendation system.

**Trip Recommendation** aims to propose a series of POIs for individuals with specific travel preferences, such as the number of attractions, starting and ending points, and so on. He et al. [99] introduced a novel context-aware POI embedding model that jointly learns the impact of POI popularity, user preferences, and co-occurring POIs. Uniquely integrating visual features, Photo2Trip [386] utilized geo-tagged photos with collaborative filtering models to enhance personalized tour recommendations. GraphTrip [77] addressed the sparsity of data and the need for more personalized trip recommendations by effectively integrating diverse knowledge domains. The paper introduces a dual-grained mobility learning approach that uses spatio-temporal graph representation.

### 5.3. Economy

The economy serves as a critical indicator of urban development, and data collected from cities inherently reflects the economic conditions of the regions. For instance, the density of POIs can indicate the popularity of an area, satellite images can provide insights into the urbanization level in a region, and human mobility data can contribute to the estimation of the economic vitality in regions.

**Real estate price** is a vital indicator for a city which is affected by various factors. To manage citywide real estate information, Du et al. [63] designed a dual-level collective learning (*DLCL*) framework to collectively learn spatial representation through intra-region and inter-region geographical structure knowledge which contain POI information as well as geographical and human mobility information for accurate prediction of the real estate price of Beijing. In their framework, intra-region POI and other side information are constructed through multi-view POI-POI networks. The constructed intra-region structure and inter-region autocorrelation are embedded through the Adversarial AutoEncoder.

Jenkins et al. [116] proposed a model named *RegionEncoder* for various data sources and conducted experiments on Chicago and New York City for the prediction of real estate prices as well. They added satellite image data to the input of the model as a supplementary and gained impressive results in the experiment.

In addition to real estate price prediction, the study conducted by Xi et al. [309] introduced an attention model that combines satellite imagery with POI information to **estimate various economic indicators**. Their methodology addresses the limitations of previous studies on satellite image data by leveraging the complementary nature of POI data in capturing human activity factors. Li et al. [154] conducted the first research to organize multi-view urban images (including satellite images and street-view images) by leveraging city structural information. The effective combination of the street-view image data contributed to a significant improvement of 10% in urban economic indicators predicting compared to previous studies before 2022. Besides, Liu et al. [186] also emphasized the necessity of integrating street-view images with satellite images in their research about urban economy.

### 5.4. Public Safety and Security

In the process of daily urban operations, ensuring public safety and security is a critical responsibility of cities and a standing focus for city administrators and researchers. The activities of humans and the movement of vehicles often encounter various safety hazards. Additionally, natural events like landslides and air pollution can also pose threats to people’s lives. The various urban big data provide us with oppor-

tunities to gain insights into unsafe factors and predict unsafe events. By integrating and analyzing diverse data sources such as human mobility, social media data, sensor data, and emergency response records, we can identify patterns and trends that may indicate potential safety risks. For example, Zhang et al. [368] proposed a multi-view region embedding framework with human mobility, region attribution, and crime records of New York City to predict the number of crime events in each region. Jiang et al. [124] designed a framework to utilize vehicle trajectories and road environment data to infer traffic violation-prone locations in Xiamen City.

**Traffic safety**, being a paramount concern for numerous countries, has emerged as a prominent focus in recent data fusion studies pertaining to urban security. Wang et al. [272] proposed a Geographical and Semantic spatio-temporal Network (*GSNet*) model for predicting traffic accidents from taxi order, POI, and weather data. They conducted an experiment on New York City and Chicago based on real-world traffic accident datasets, the result demonstrated a satisfactory result for their model. Then Wang et al. [288] designed a cross-scale feature fusion mechanism to integrate features from different scale data as well as a feature fusion component to integrate multi-view features. Based on this work, a multi-task learning framework that can simultaneously predict fine and coarse-grained traffic accident risks was proposed.

**Environment safety** and the mitigation of natural disasters have also been remarkable research focal points. These areas of study have garnered significant attention due to their profound implications for urban environments and the well-being of communities. Researchers have made notable strides in leveraging data fusion techniques to enhance environmental monitoring, early warning systems, and disaster response strategies. For instance, based on the existing study of trajectories, Luo et al. [188] combined traffic trajectory information with road networks to efficiently identify traffic bottlenecks on the road. Chen et al. [39] and You et al. [342] focused on studying the road obstacle situation or road risks after disasters in a city from its vehicle trajectory, satellite image, and meteorology data. Besides, Song et al. [249] built an intelligent system named *Deep-Mob* in Japan based on users' trajectories, earthquake records, text reports, and transportation network data to analyze and predict human behavior and mobility following natural disasters.

### 5.5. Social

The advances in wireless communication and location acquisition technologies enable people to add a location dimension to traditional social networks [204]. With the advancement of data fusion advantage and the popularity of recommendation devices, services based on Location-based Social Networks (LBSN) have penetrated people's lives. Users on social media generate a variety of content with geo-information every day. Such posts combine geographical coordinates (always in GPS location) and user-generated content (text, image, or audio) which might be associated with the semantic meaning of those places. It is strongly promising to fusion the geographical information and other information through the users' posts to bridge the gap between users' activities in digital and physical

worlds and contribute to various downstream tasks such as POI or friend recommendation [42, 31, 219, 69, 155] and community analysis and detection [256, 184, 316].

**Recommendation on LBSN** based on deep learning is a complex task in real applications as the irregular spatial structure of geo-location data is non-Euclidean and traditional neural network-based deep learning is incapable of such non-Euclidean data. So graph neural network is widely used to understand the spatial structure information with geo-location data. Fang et al. [69] proposed a multi-graph fusion approach for POI recommendation by constructing a user-POI interaction graph on LBSN. For LBSN friend recommendation, Li et al. [155] designed a learning framework by leveraging multi-graph to model raw LBSN data and defining various connections between nodes to represent spatio-temporal information. In their methodology, a contrastive learning model was proposed to integrate spatio-temporal features of human trajectories in cities for user node embedding learning.

### 5.6. Environment

Rapid urbanization may result in a potential threat to cities' environment. For example, it is reported that ninety percent of air pollution is caused by urban transportation. Environment protection is an essential topic for urban computing. We have witnessed much research on the environment from different aspects of urban computing, such as air quality prediction [138, 114, 129, 279, 166], noise controlling [1, 117, 118, 64, 185] and weather forecasting [399, 341, 36, 247, 87]. Various data sources cloud provide information from various aspects for us to sense the environment. Further, the fusion those information from multiple sources was proven to be effective in understanding and predicting the variation of the environment in a city.

Zheng et al. [399] conducted research on cities' air quality based on multi-source big data. They proposed a multi-view hybrid model to predict future 48-hour air quality of a single station point by converging historical air quality data of one station, current meteorological data in the area, weather forecasting data as well as air quality data from other stations. The various meteorological factors and spatial relations between different stations are both considered by effectively fusing these data which significantly improved the capability and accuracy of fine-grain air quality prediction. Further, Ma et al. [192] utilized a graph neural network to model spatio-temporal correlations between different meteorological variables and different stations as graph neural networks have demonstrated excellent performances for modeling spatio-temporal information. This kind of research makes city-level fine-grained weather forecasting possible with multi-source spatio-temporal data.

### 5.7. Energy

From the perspective of energy, cities can be seen as machines that consume a significant amount of energy. With the acceleration of urbanization and the growing global energy concerns, technologies that enable the perception of energy consumption and energy efficiency have become increasingly important. In urban computing, the fusion of diverse data sources

allows us to have a more comprehensive understanding of urban energy consumption and make more efficient urban planning or transportation decisions to help reduce energy consumption.

### 5.7.1. Energy Efficient Urban Planning

Electric vehicles are regarded as a promising solution for green transportation and sustainable cities. With the gradual expansion of the market share of electrical vehicles, the demand for car charging facilities is increasing. One important issue in urban planning is the selection of charging station locations. Proper selection of these locations not only optimizes city management but also meets the daily commuting needs of users, reducing unnecessary energy consumption during the search for charging stations. The selection of charging station locations should take into account various factors, such as the economic situation of the region, the number of electric vehicles in circulation, residents' travel patterns, road conditions, and traffic conditions. This necessitates the integration and analysis of multiple data sources to estimate the regional charging demand and identify suitable locations for charging stations. Tu et al. [267] proposed a spatio-temporal demand coverage location model based on the taxi GPS data in Shenzhen, China as well as data from the charging stations. This model provided a charging station sitting approach by considering the waiting time which could hugely reduce energy consumption for charging. He et al. [100] developed a location method with the consideration of cars' driving range extracted from the GPS data. Battaia et al. [16] proposed a framework for charging situation sitting considering the real-life problem of charging infrastructure and sustainability.

### 5.7.2. Energy Efficient Urban Transportation

Transportation energy efficiency aims to recommend and navigate energy-efficient travel routes for every transportation participant. Nowadays, within the worldwide energy shortage, fuel prices have been continuously rising [387]. However, traffic congestion in 439 cities in the United States resulted in a total loss of 3.3 billion gallons of wasted fuel [237] in only one year. Traditional route planning studies primarily emphasized faster and more reliable routes, often overlooking the differences in energy consumption among different travel routes and modes for the same travel demand. For instance, in certain situations, taking a route through a congested city center may result in faster arrival times compared to taking a detour on a highway. Despite the first route being shorter in distance, the driver may experience delays due to frequent stops and slow speeds caused by city congestion.

In application, recommending energy-efficient routes for users is a highly complex task. Unlike traditional shortest path problems, it requires considering various factors such as vehicle efficiency, driving habits, road conditions, POIs, and more. The key to addressing this task lies in effectively integrating information from multiple sources. Wang et al. [293] researched personalized fuel-efficient route recommendations based on history trajectories. The temporal information from the trajectories as well as the driving factors are input into a transformer model to compose fuel efficiency prediction on the target route. By

combining with a genetic algorithm for the best recommendation, this research could effectively find the fuel-efficient routes on the real-world dataset. DeepFEC [66] proposed a framework to predict road-level energy consumption through a fusion of contextual vehicle data such as type, weight and engine configuration, and spatio-temporal traffic data on roads. In this framework, a deep convolution residual model is designed to capture the spatial dynamic information from the data and a Bi-LSTM model is responsible for the temporal information. Oh et al. [209] is the first large dataset that combines trajectories of personal cars with users' information, making it possible to mine users' driving behaviors. Based on this multi-source dataset, Lai et al. [145] designed a meta-learning framework for predicting vehicle energy consumption in Ann Arbor, Michigan, USA that can be fine-tuned based on a user's driving preference.

## 6. Challenges and Future Directions

While urban multi-modal research has made significant advancements in recent years, several challenging issues persist, highlighting directions for potential future research. As shown in Figure 27, we summarize these challenges and suggest potentially feasible research directions as follows:



Figure 27: Challenges and future directions of cross-domain data fusion in Urban Computing.

- **LLM-enhanced Application:** Since the advent of GPT-4 [211] and Sora [22], the academic community has embarked on extensive research on the roles of LLM as predictors [169, 106], enhancers [336, 221], controllers [73, 242, 317], and evaluators [255, 169]. However, the field of cross-domain data fusion in urban computing is still in its initial stage in terms of exploring how to apply LLMs effectively. For example, there exists research [232, 24, 195] focusing on LLM as urban predictors but also as naive reasoners yet; whereas others [105, 141] attempt to leverage LLM as model backbone but such works are limited within remote sensing domain instead of general urban computing. A critical limitation currently facing the application of LLMs in Urban Computing is their inability to perceive spatial relationships and dependencies inherent in natural spaces [408], a trait typically associated with language models. For instance, mainstream LLMs struggle to comprehend intricate spatial orientations and positions of urban entities. Moreover, the majority of publicly accessible LLMs are predominantly text-based and exhibit limited performance in other modalities. Nevertheless, we are encouraged by the growing number of successful initiatives in the graph learning domain that aim to enhance LLMs' comprehension of complex relations within

graph data [259, 160]. Furthermore, the latest multimodal LLMs, such as GPT-4o, demonstrate considerable promise in offering effective multimodal solutions. Therefore, we eagerly anticipate that the research community will focus on various applications of LLMs in urban computing, exploring their impact similar to their roles in NLP and other domains.

- **Agent-based Simulation:** Agent-based simulation aims to model the behavior of individual components or agents to comprehend their interactions and how they collectively contribute to the functioning of the entire system [310, 280]. It has been extensively used in various fields such as biology [6], ecology [89], sociology [72], etc., to model systems where individual entities influence collective behavior. However, the early attempts at agent-based simulation [70, 81, 389] are limited because they are not autonomous and need human-defined rules or goals, which may not simulate the complex dynamics of urban systems. Hence, LLM-driven agents such as Urban Generative Intelligence (UGI) [317], can serve as up-to-date solutions for simulating urban dynamics based on cross-domain urban data. This paradigm not only propels forward the field of urban computing, but also paves the way for future cities that are more adaptive and responsive to the evolving needs of their inhabitants.
- **Multi-modal Causal Learning:** Causal learning or inference aims to investigate causal relationships between variables, ensuring stable and robust learning and inference [206, 220]. Integrating deep learning techniques with causal inference has shown great success in recent years, especially in the fields of spatio-temporal graph forecasting [403, 312, 149], CV [290, 359, 168], NLP [269, 371, 264], and recommender systems [393, 75]. However, regarding cross-domain data fusion in urban computing, the application of causal learning is still in its early stages. One of the most severe challenges is to represent the complex cross-modal causality in urban scenarios. The deep learning community has yet to converge on a universally accepted approach for accurately yet efficiently representing cross-modal relations. However, as an increasing number of studies delve into the semantic essence of information across various modalities and advance the fields of graph learning, representation learning, and urban foundational models, we eagerly anticipate further research on multi-modal causal learning of urban data, aiming to improve the interpretability of intricate and dynamic urban systems.
- **Multi-source Data Privacy:** Urban data can be highly sensitive, especially in multi-source scenarios in the field of economy and healthcare. When models are trained upon such shared data, there is a risk that they may memorize specific information from the training data, potentially compromising the privacy of individuals. Particularly in the context of deploying data fusion models, the inclusion of data from diverse sources and modalities can indeed bolster the models' performance. However, this also heightens the risk of inadvertently exposing sensitive privacy data. Given these privacy concerns, many institutions and city governments are

reluctant to share urban data, thereby impeding the deployment of deep learning models. This poses a significant barrier to the advancement of the community. Consequently, there is a pressing need for research that explores the integration of privacy-preserving techniques, such as differential privacy [65, 326] and federated learning [357, 153]. The goal is to protect data privacy while still reaping the benefits of cross-domain data fusion in urban computing.

- **Open Benchmark:** The challenge of developing an open benchmark for cross-domain data fusion in urban computing lies in the complexity of integrating diverse data sources, such as sensor data, images, and even text, to understand urban environments comprehensively. This complexity arises due to the heterogeneity of data formats, modalities, and the need for effective fusion methods to extract meaningful insights. A potential solution involves collaborative efforts to standardize data formats, develop unified evaluation metrics, and establish shared benchmarks that facilitate the evaluation and comparison of cross-domain data fusion models.
- **Downstream Task Diversity:** Existing urban research predominantly concentrates on specific task domains such as transportation and urban planning, overlooking the inherent diversity of challenges in real-life urban environments. This limitation exists due to the compartmentalized nature of current research efforts, hindering a holistic understanding of cross-domain data fusion in urban computing. Therefore, we anticipate a more extensive scope of urban research encompassing diverse applications in the realms of economy, society, and environment, providing a thorough comprehension of the intricate conditions prevailing in urban settings.
- **Computation Efficiency:** Current urban research emphasizes the fulfillment of specific applications within cross-domain urban computing, but it overlooks the crucial aspect of computational efficiency. This oversight hampers the practical deployment of these computational models in real-life scenarios. Addressing this challenge requires our focus on optimizing computation efficiency, involving model compression (e.g., knowledge distillation [90], low-rank decomposition [331, 159], and quantization [58, 59]), efficient training (e.g., prompt tuning [180] and hardware-assisted attention acceleration [55, 54]), and efficient architecture (e.g., mixture of experts [44, 241]), to enhance the feasibility and effectiveness of deploying cross-domain data fusion solutions in practical urban environments.

## 7. Conclusion

In this survey, we present an extensive and up-to-date survey of cross-domain data fusion tailored for urban computing, aiming to offer a fresh perspective on this evolving field by introducing a novel taxonomy that categorizes the reviewed fusion methods. In particular, we initially explore the data perspective to understand the significance of each modality and data source, classify the methodology into four types (i.e., *feature-based*, *alignment-based*, *contrast-based*, and *generation-based*), and

further categorize cross-domain urban applications into seven types. Besides, We succinctly outline the possible challenges while shedding light on promising directions for future research. The potential for urban computing-related investigations within this captivating cross-domain fusion field is limitless. We hope this can ignite more curiosity and cultivate a long-lasting enthusiasm for cross-domain urban studies, therefore achieving real urban intelligence.

## Appendix A. Github Link

Please refer to <https://github.com/yoshall/Awesome-Multimodal-Urban-Computing> for comprehensive and up-to-date paper list.

## References

- [1] Abbaspour, M., Karimi, E., Nassiri, P., Monazzam, M.R., Taghavi, L., 2015. Hierachal assessment of noise pollution in urban areas—a case study. *Transportation Research Part D: Transport and Environment* 34, 95–103.
- [2] Afyouni, I., Al Aghbari, Z., Razack, R.A., 2022. Multi-feature, multi-modal, and multi-source social event detection: A comprehensive survey. *Information Fusion* 79, 279–308.
- [3] Aheto, J.M.K., Olowe, I.D., Chan, H.M.T., Ekeh, A., Dieng, B., Fafunmi, B., Setayesh, H., Atuhaire, B., Crawford, J., Tatem, A.J., Utazi, C.E., 2023. Geospatial analyses of recent household surveys to assess changes in the distribution of zero-dose children and their associated factors before and during the covid-19 pandemic in nigeria. *Vaccines* 11. URL: <https://doi.org/10.3390/vaccines11121830>, doi:[10.3390/vaccines11121830](https://doi.org/10.3390/vaccines11121830).
- [4] Alfarrarjeh, A., Yang, X., Jabal, A.A., Kim, S.H., Shahabi, C., 2021. Exploring the spatial-visual locality of geo-tagged urban street images, in: 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE. pp. 104–110.
- [5] Amirian, J., Van Toll, W., Hayet, J.B., Pettré, J., 2019. Data-driven crowd simulation with generative adversarial networks, in: Proceedings of the 32nd International Conference on Computer Animation and Social Agents, pp. 7–10.
- [6] An, G., 2021. Agent-based modeling in translational systems biology. *Complex Systems and Computational Biology Approaches to Acute Inflammation: A Framework for Model-based Precision Medicine* , 31–52.
- [7] Anbalagan, B., 2022. Event location detection from online clustering algorithms using geo-tagged user data in social streams, in: *Disruptive Technologies for Big Data and Cloud Applications: Proceedings of ICB-DCC 2021*. Springer, pp. 227–235.
- [8] Anderegg, W.R., Trugman, A.T., Badgley, G., Anderson, C.M., Bartuska, A., Ciais, P., Cullenward, D., Field, C.B., Freeman, J., Goetz, S.J., et al., 2020. Climate-driven risks to the climate mitigation potential of forests. *Science* 368, eaaz7005.
- [9] Arslan Ay, S., Zhang, L., Kim, S.H., He, M., Zimmermann, R., 2009. Grvs: a georeferenced video search engine, in: *Proceedings of the 17th ACM International Conference on Multimedia*, pp. 977–978.
- [10] Bai, L., Huang, W., Zhang, X., Du, S., Cong, G., Wang, H., Liu, B., 2023. Geographic mapping with unsupervised multi-modal representation learning from VHR images and POIs. *ISPRS Journal of Photogrammetry and Remote Sensing* 201, 193–208. doi:[10.1016/j.isprsjprs.2023.05.006](https://doi.org/10.1016/j.isprsjprs.2023.05.006).
- [11] Bai, L., Yao, L., Kanhere, S., Wang, X., Sheng, Q., et al., 2019a. Stg2seq: Spatial-temporal graph to sequence model for multi-step passenger demand forecasting. *arXiv preprint arXiv:1905.10069* .
- [12] Bai, L., Yao, L., Kanhere, S.S., Wang, X., Liu, W., Yang, Z., 2019b. Spatio-temporal graph convolutional and recurrent networks for city-wide passenger demand prediction, in: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 2293–2296.
- [13] Balsebre, P., Yao, D., Cong, G., Hai, Z., 2022. Geospatial entity resolution, in: *Proceedings of the ACM Web Conference 2022*, Association for Computing Machinery, New York, NY, USA. pp. 3061–3070. doi:[10.1145/3485447.3512026](https://doi.org/10.1145/3485447.3512026).
- [14] Baltrušaitis, T., Ahuja, C., Morency, L.P., 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 423–443.
- [15] Barbosa, H., Barthelemy, M., Ghoshal, G., James, C.R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J.J., Simini, F., Tomasini, M., 2018. Human mobility: Models and applications. *Physics Reports* 734, 1–74.
- [16] Battaïa, O., Dolgui, A., Guschinsky, N., Rozin, B., 2023. Milp model for fleet and charging infrastructure decisions for fast-charging city electric bus services. *Computers & Industrial Engineering* , 109336.
- [17] Bellemans, T., De Schutter, B., De Moor, B., 2002. Models for traffic control. *JOURNAL A* 43, 13–22.
- [18] Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and gis: A review. *Landscape and Urban Planning* 215, 104217.
- [19] Bin, J., Gardiner, B., Liu, H., Li, E., Liu, Z., 2023. Rhpmf: A context-aware matrix factorization approach for understanding regional real estate market. *Information Fusion* 94, 229–242.
- [20] Bing, J., Chen, M., Yang, M., Huang, W., Gong, Y., Nie, L., 2023. Pre-trained semantic embeddings for POI categories based on multiple contexts. *IEEE Transactions on Knowledge and Data Engineering* 35, 8893–8904. doi:[10.1109/TKDE.2022.3218851](https://doi.org/10.1109/TKDE.2022.3218851).
- [21] Breunig, M., Bradley, P.E., Jahn, M., Kuper, P., Mazroob, N., Rösch, N., Al-Door, M., Stefanakis, E., Jadidi, M., 2020. Geospatial data management research: Progress and future directions. *ISPRS International Journal of Geo-Information* 9, 95.
- [22] Brooks, T., Peebles, B., Homes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A., 2024. Video generation models as world simulators URL: <https://openai.com/research/video-generation-models-as-world-simulators>.
- [23] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33, 1877–1901.
- [24] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kammar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al., 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* .
- [25] Bui, T.H., 2023. Automatic construction of POI address lists at city streets from geo-tagged photos and web data: a case study of San Jose City. *Multimedia Tools and Applications* , 1–22.
- [26] Burke, M., Driscoll, A., Lobell, D.B., Ermon, S., 2021. Using satellite imagery to understand and promote sustainable development. *Science* 371, eabe8628.
- [27] Bwire, H., Zengo, E., 2020. Comparison of efficiency between public and private transport modes using excess commuting: An experience in dar es salaam. *Journal of Transport Geography* 82, 102616.
- [28] Cai, T., Wan, H., Wu, F., Wen, H., Guo, S., Wu, L., Hu, H., Lin, Y., 2023. M 2 g4rtpp: A multi-level and multi-task graph model for instant-logistics route and time joint prediction, in: *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, IEEE. pp. 3296–3308.
- [29] Cao, D., Jia, F., Arik, S.O., Pfister, T., Zheng, Y., Ye, W., Liu, Y., 2023. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. *arXiv preprint arXiv:2310.04948* .
- [30] Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.A., Li, S.Z., 2022. A survey on generative diffusion model. *arXiv preprint arXiv:2209.02646* .
- [31] Cao, K., Guo, J., Meng, G., Liu, H., Liu, Y., Li, G., 2020. Points-of-interest recommendation algorithm based on LBSN in edge computing environment. *IEEE Access* 8, 47973–47983. doi:[10.1109/ACCESS.2020.2979922](https://doi.org/10.1109/ACCESS.2020.2979922).
- [32] Carozzi, F., Roth, S., 2023. Dirty density: Air quality and the density of american cities. *Journal of Environmental Economics and Management* 118, 102767.
- [33] Chandra, D.K., Leopold, J., Fu, Y., 2021. NodeSense2Vec: Spatiotemporal context-aware network embedding for heterogeneous urban mobility data, in: *2021 IEEE International Conference on Big Data (Big Data)*, pp. 2884–2893. doi:[10.1109/BigData52589.2021.9672072](https://doi.org/10.1109/BigData52589.2021.9672072).

- [34] Chang, C., Peng, W.C., Chen, T.F., 2023. Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms. arXiv preprint arXiv:2308.08469 .
- [35] Chen, B., Xu, B., Gong, P., 2021a. Mapping essential urban land use categories (euluc) using geospatial big data: Progress, challenges, and opportunities. *Big Earth Data* 5, 410–441.
- [36] Chen, G., Liu, S., Jiang, F., 2022a. Daily weather forecasting based on deep learning model: A case study of shenzhen city, china. *Atmosphere* 13, 1208.
- [37] Chen, G., Su, Y., Zhang, X., Hu, A., Chen, G., Feng, S., Xiang, J., Zhang, J., Zheng, Y., 2022b. A cross-city federated transfer learning framework: A case study on urban region profiling. doi:[10.48550/arXiv.2206.00007](https://doi.org/10.48550/arXiv.2206.00007), arXiv:[2206.00007](https://arxiv.org/abs/2206.00007).
- [38] Chen, J., Zhou, C., Li, F., 2020. Quantifying the green view indicator for assessing urban greening quality: An analysis based on internet-crawling street view data. *Ecological Indicators* 113, 106192.
- [39] Chen, L., Fan, X., Wang, L., Zhang, D., Yu, Z., Li, J., Thi Mai Trang, N., Pan, G., Wang, C., 2018. RADAR: Road obstacle identification for disaster response leveraging cross-domain urban data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 1–23. doi:[10.1145/3161159](https://doi.org/10.1145/3161159).
- [40] Chen, L., Lu, C., Yuan, F., Jiang, Z., Wang, L., Zhang, D., Luo, R., Fan, X., Wang, C., 2021b. UVLens: Urban village boundary identification and population estimation leveraging open government data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 57:1–57:26. doi:[10.1145/3463495](https://doi.org/10.1145/3463495).
- [41] Chen, W., Wang, G., Zeng, J., 2023. Impact of urbanization on ecosystem health in chinese urban agglomerations. *Environmental Impact Assessment Review* 98, 106964.
- [42] Chen, X., Zeng, Y., Cong, G., Qin, S., Xiang, Y., Dai, Y., 2015. On information coverage for location category based point-of-interest recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence* 29. doi:[10.1609/aaai.v29i1.9191](https://doi.org/10.1609/aaai.v29i1.9191).
- [43] Chen, Z., Chen, L., Cong, G., Jensen, C.S., 2021c. Location-and keyword-based querying of geo-textual data: a survey. *The VLDB Journal* 30, 603–640. URL: <https://doi.org/10.1007/s00778-021-00661-w>, doi:[10.1007/s00778-021-00661-w](https://doi.org/10.1007/s00778-021-00661-w).
- [44] Chen, Z., Deng, Y., Wu, Y., Gu, Q., Li, Y., 2022c. Towards understanding the mixture-of-experts layer in deep learning, in: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (Eds.), *Advances in Neural Information Processing Systems*, Curran Associates, Inc.. pp. 23049–23062. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf).
- [45] Cheng, Q., Jing, Q., Collender, P.A., Head, J.R., Li, Q., Yu, H., Li, Z., Ju, Y., Chen, T., Wang, P., Cleary, E., Lai, S., 2023. Prior water availability modifies the effect of heavy rainfall on dengue transmission: a time series analysis of passive surveillance data from southern china. *Frontiers in Public Health* URL: <https://doi.org/10.3389/fpubh.2023.1287678>, doi:[10.3389/fpubh.2023.1287678](https://doi.org/10.3389/fpubh.2023.1287678).
- [46] Chengchuang, L., Chun, S., Gansen, Z., et al., 2021. Review of image data augmentation in computer vision. *Journal of Frontiers of Computer Science & Technology* 15, 583.
- [47] Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y., 2009. Nus-wide: A real-world web image database from national university of singapore, in: *ACM International Conference on Image and Video Retrieval*, pp. 48:1–48:9.
- [48] CondorFerries, 2023. Explore solo travel trends & stats by demographics, destination, industry & why solo travel continues to rise! <https://www.condorferries.co.uk/solo-travel-statistics/>.
- [49] Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., Ermon, S., 2022. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems* 35, 197–211.
- [50] Conger, K., 2023. So what do we call twitter now anyway? The New York Times Archived from the original on October 12, 2023. Retrieved August 29, 2023.
- [51] Cui, J., Li, Z., Yan, Y., Chen, B., Yuan, L., 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. arXiv preprint arXiv:2306.16092 .
- [52] Dai, J., Yang, B., Guo, C., Ding, Z., 2015. Personalized route recommendation using big trajectory data, in: *2015 IEEE 31st international conference on data engineering*, IEEE. pp. 543–554.
- [53] Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S., 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 .
- [54] Dao, T., 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691 .
- [55] Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C., 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems* 35, 16344–16359.
- [56] Deldari, S., Xue, H., Saeed, A., He, J., Smith, D.V., Salim, F.D., 2022. Beyond just vision: A review on self-supervised representation learning on multimodal and temporal data. arXiv preprint arXiv:2206.02353 .
- [57] Desai, K., Johnson, J., 2021. Virtex: Learning visual representations from textual annotations, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11162–11173.
- [58] Dettmers, T., Lewis, M., Belkada, Y., Zettlemoyer, L., 2022. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. arXiv preprint arXiv:2208.07339 .
- [59] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L., 2023. Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314 .
- [60] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .
- [61] Ding, R., Chen, B., Xie, P., Huang, F., Li, X., Zhang, Q., Xu, Y., 2023. Mgeo: Multi-modal geographic language model pre-training, in: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 185–194.
- [62] Doi, K., 2015. Cities and transportation. *Traffic and Safety Sciences—Interdisciplinary Wisdom of IATSS*, International Association of Traffic and Safety Sciences , 12–21.
- [63] Du, J., Zhang, Y., Wang, P., Leopold, J., Fu, Y., 2019. Beyond geo-first law: Learning spatial representations via integrated autocorrelations and complementarity, in: *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 160–169. doi:[10.1109/ICDM.2019.00026](https://doi.org/10.1109/ICDM.2019.00026).
- [64] Dutta, J., Pramanick, P., Roy, S., 2017. Noisesense: Crowdsourced context aware sensing for real time noise pollution monitoring of the city, in: *2017 IEEE international conference on advanced networks and telecommunications systems (ANTS)*, IEEE. pp. 1–6.
- [65] Dwork, C., 2008. Differential privacy: A survey of results, in: *International conference on theory and applications of models of computation*, Springer. pp. 1–19.
- [66] Elmi, S., Tan, K.L., 2021. Deepfec: energy consumption prediction under real-world driving conditions for smart cities, in: *Proceedings of the Web Conference 2021*, pp. 1880–1890.
- [67] Fadhel, M.A., Duham, A.M., Saihood, A., Sewify, A., Al-Hamadani, M.N., Albahri, A., Alzubaidi, L., Gupta, A., Mirjalili, S., Gu, Y., 2024. Comprehensive systematic review of information fusion methods in smart cities and urban environments. *Information Fusion* , 102317.
- [68] Fan, Z., Zhang, F., Loo, B.P.Y., Ratti, C., 2023. Urban visual intelligence: Uncovering hidden city profiles with street view images. *Proceedings of the National Academy of Sciences* 120, e2220417120. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.2220417120>, doi:[10.1073/pnas.2220417120](https://doi.org/10.1073/pnas.2220417120), arXiv:<https://www.pnas.org/doi/pdf/10.1073/pnas.2220417120>.
- [69] Fang, J., Meng, X., Qi, X., 2023. A top-k POI recommendation approach based on LBSN and multi-graph fusion. *Neurocomputing* 518, 219–230. doi:[10.1016/j.neucom.2022.10.048](https://doi.org/10.1016/j.neucom.2022.10.048).
- [70] Farmer, J.D., Foley, D., 2009. The economy needs agent-based modelling. *Nature* 460, 685–686.
- [71] Fisch, D., Kalkowski, E., Sick, B., 2013. Knowledge fusion for probabilistic generative classifiers with data mining applications. *IEEE Transactions on Knowledge and Data Engineering* 26, 652–666.
- [72] Flache, A., Mäis, M., Keijzer, M.A., 2022. Computational approaches in rigorous sociology: agent-based computational modeling and computational social science. *Handbook of Sociological Science* , 57–72.
- [73] Foosherian, M., Purwins, H., Rathnayake, P., Alam, T., Teimao, R., Thoben, K.D., 2023. Enhancing pipeline-based conversational agents with large language models. arXiv preprint arXiv:2309.03748 .
- [74] Fu, Y., Wang, P., Du, J., Wu, L., Li, X., 2019. Efficient region embedding

- with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 906–913.
- [75] Gao, C., Zheng, Y., Wang, W., Feng, F., He, X., Li, Y., 2022a. Causal inference in recommender systems: A survey and future directions. ACM Transactions on Information Systems .
- [76] Gao, N., Xue, H., Shao, W., Zhao, S., Qin, K.K., Prabowo, A., Rahaman, M.S., Salim, F.D., 2022b. Generative adversarial networks for spatio-temporal data: A survey. ACM Transactions on Intelligent Systems and Technology (TIST) 13, 1–25.
- [77] Gao, Q., Wang, W., Huang, L., Yang, X., Li, T., Fujita, H., 2023a. Dual-grained human mobility learning for location-aware trip recommendation with spatial-temporal graph knowledge fusion. Information Fusion 92, 46–63.
- [78] Gao, Q., Wang, W., Huang, L., Yang, X., Li, T., Fujita, H., 2023b. Dual-grained human mobility learning for location-aware trip recommendation with spatial-temporal graph knowledge fusion. Information Fusion 92, 46–63. doi:[10.1016/j.inffus.2022.11.018](https://doi.org/10.1016/j.inffus.2022.11.018).
- [79] Gao, Q., Zhou, F., Zhong, T., Trajcevski, G., Yang, X., Li, T., 2022c. Contextual spatio-temporal graph representation learning for reinforced human mobility mining. Information Sciences 606, 230–249.
- [80] Gao, R., Chen, K., Xie, E., Hong, L., Li, Z., Yeung, D.Y., Xu, Q., 2023c. Magicdrive: Street view generation with diverse 3d geometry control. arXiv preprint arXiv:2310.02601 .
- [81] Geanakoplos, J., 2010. The leverage cycle. NBER macroeconomics annual 24, 1–66.
- [82] Geng, X., Li, Y., Wang, L., Zhang, L., Yang, Q., Ye, J., Liu, Y., 2019a. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting, in: Proceedings of the AAAI conference on artificial intelligence, pp. 3656–3663.
- [83] Geng, X., Liu, H., Lee, L., Schuurmans, D., Levine, S., Abbeel, P., 2022. Multimodal masked autoencoders learn transferable representations. arXiv preprint arXiv:2205.14204 .
- [84] Geng, X., Wu, X., Zhang, L., Yang, Q., Liu, Y., Ye, J., 2019b. Multi-modal graph interaction for multi-graph convolution network in urban spatiotemporal forecasting. doi:[10.48550/arXiv.1905.11395](https://doi.org/10.48550/arXiv.1905.11395), arXiv:1905.11395.
- [85] GeoVid Project, . GeoVid Project. URL: <http://geovid.org/>.
- [86] Gerla, M., Kleinrock, L., 1980. Flow control: A comparative survey. IEEE Transactions on Communications 28, 553–574.
- [87] Ghoneim, O.A., Manjunatha, B., et al., 2017. Forecasting of ozone concentration in smart city using deep learning, in: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE. pp. 1320–1326.
- [88] Gong, Z., Ma, Q., Kan, C., Qi, Q., 2019. Classifying street spaces with street view images for a spatial indicator of urban functions. Sustainability 11, 6424.
- [89] González-Crespo, C., Martínez-López, B., Conejero, C., Castillo-Contreras, R., Serrano, E., López-Martín, J.M., Lavín, S., López-Olvera, J.R., 2023. Predicting human-wildlife interaction in urban environments through agent-based models. Landscape and Urban Planning 240, 10478.
- [90] Gou, J., Yu, B., Maybank, S.J., Tao, D., 2021. Knowledge distillation: A survey. International Journal of Computer Vision 129, 1789–1819.
- [91] Guo, C., Yang, B., Hu, J., Jensen, C.S., Chen, L., 2020. Context-aware, preference-based vehicle routing. The VLDB Journal 29, 1149–1170.
- [92] Guo, J., Gong, Z., 2016. A nonparametric model for event discovery in the geospatial-temporal space, in: Proceedings of the 25th ACM International Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA. pp. 499–508. doi:[10.1145/2983323.2983790](https://doi.org/10.1145/2983323.2983790).
- [93] Guo, S., Chen, C., Wang, J., Ding, Y., Liu, Y., Xu, K., Yu, Z., Zhang, D., 2022. A force-directed approach to seeking route recommendation in ride-on-demand service using multi-source urban data. IEEE Transactions on Mobile Computing 21, 1909–1926. doi:[10.1109/TMC.2020.3033274](https://doi.org/10.1109/TMC.2020.3033274).
- [94] Guo, S., Chen, C., Wang, J., Liu, Y., Xu, K., Yu, Z., Zhang, D., Chiu, D.M., 2019. Rod-revenue: Seeking strategies analysis and revenue prediction in ride-on-demand service using multi-source urban data. IEEE Transactions on Mobile Computing 19, 2202–2220.
- [95] Hajela, G., Chawla, M., Rasool, A., 2021. A multi-dimensional crime spatial pattern analysis and prediction model based on classification. ETRI Journal 43, 272–287.
- [96] Han, P., Wang, J., Yao, D., Shang, S., Zhang, X., 2021. A graph-based approach for trajectory similarity computation in spatial networks, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 556–564.
- [97] Hao, X., Chen, W., Yan, Y., Zhong, S., Wang, K., Wen, Q., Liang, Y., 2024. Urbanlp: A multi-granularity vision-language pre-trained foundation model for urban indicator prediction. arXiv preprint arXiv:2403.16831 .
- [98] Hashem, I.A.T., Usmani, R.S.A., Almutairi, M.S., Ibrahim, A.O., Zakari, A., Alotaibi, F., Alhashmi, S.M., Chiroma, H., 2023. Urban computing for sustainable smart cities: Recent advances, taxonomy, and open research challenges. Sustainability 15, 3916.
- [99] He, J., Qi, J., Ramamohanarao, K., 2019. A joint context-aware embedding for trip recommendations, in: 2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE. pp. 292–303.
- [100] He, J., Yang, H., Tang, T.Q., Huang, H.J., 2018. An optimal charging station location model with the consideration of electric vehicle's driving range. Transportation Research Part C: Emerging Technologies 86, 641–654.
- [101] He, J., Yang, K., Tang, W., Lu, H., Qin, J., Chen, Y., Li, X., 2020. The first high-resolution meteorological forcing dataset for land process studies over China. Scientific Data 7, 25.
- [102] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16000–16009.
- [103] Hermans, M., Schrauwen, B., 2013. Training and analysing deep recurrent neural networks. Advances in neural information processing systems 26.
- [104] Hou, Z., Liu, X., Cen, Y., Dong, Y., Yang, H., Wang, C., Tang, J., 2022. Graphmae: Self-supervised masked graph autoencoders, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 594–604.
- [105] Hu, Y., Yuan, J., Wen, C., Lu, X., Li, X., 2023a. Rsgpt: A remote sensing vision language model and benchmark. arXiv preprint arXiv:2307.15266 .
- [106] Hu, Z., Feng, Y., Luu, A.T., Hooi, B., Lipani, A., 2023b. Unlocking the potential of user feedback: Leveraging large language model as user simulator to enhance dialogue system. arXiv preprint arXiv:2306.09821 .
- [107] Huang, C., Wang, D., 2016. Exploiting spatial-temporal-social constraints for localness inference using online social media, in: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 287–294. doi:[10.1109/ASONAM.2016.7752247](https://doi.org/10.1109/ASONAM.2016.7752247).
- [108] Huang, C., Zhang, J., Zheng, Y., Chawla, N.V., 2018. Deepcrime: Attentive hierarchical recurrent networks for crime prediction, in: Proceedings of the 27th ACM international conference on information and knowledge management, pp. 1423–1432.
- [109] Huang, J., Chai, J., Cho, S., 2020. Deep learning in finance and banking: A literature review and classification. Frontiers of Business Research in China 14, 1–24.
- [110] Huang, J., Wang, H., Sun, Y., Shi, Y., Huang, Z., Zhuo, A., Feng, S., 2022a. ERNIE-GeoL: A geography-and-language pre-trained model and its applications in baidu maps, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3029–3039. doi:[10.1145/3534678.3539021](https://doi.org/10.1145/3534678.3539021), arXiv:2203.09127.
- [111] Huang, T., Fu, R., Chen, Y., Sun, Q., 2022b. Real-time driver behavior detection based on deep deformable inverted residual network with an attention mechanism for human-vehicle co-driving system. IEEE Transactions on Vehicular Technology 71, 12475–12488.
- [112] Huang, Y., Zhang, F., Gao, Y., Tu, W., Duarte, F., Ratti, C., Guo, D., Liu, Y., 2023a. Comprehensive urban space representation with varying numbers of street-level images. Computers, Environment and Urban Systems 106, 102043. URL: <https://www.sciencedirect.com/science/article/pii/S0198971523001060>, doi:<https://doi.org/10.1016/j.compenurbysys.2023.102043>.
- [113] Huang, Y., Zhang, F., Gao, Y., Tu, W., Duarte, F., Ratti, C., Guo, D., Liu, Y., 2023b. Comprehensive urban space representation with varying num-

- bers of street-level images. Computers, Environment and Urban Systems 106, 102043. doi:[10.1016/j.compenvurbsys.2023.102043](https://doi.org/10.1016/j.compenvurbsys.2023.102043).
- [114] Iskandaryan, D., Ramos, F., Trilles, S., 2020. Air quality prediction in smart cities using machine learning technologies based on sensor data: a review. Applied Sciences 10, 2401.
- [115] Jain, A., Zamir, A.R., Savarese, S., Saxena, A., 2016. Structural-rnn: Deep learning on spatio-temporal graphs, in: Proceedings of the ieee conference on computer vision and pattern recognition, pp. 5308–5317.
- [116] Jenkins, P., Farag, A., Wang, S., Li, Z., 2019. Unsupervised representation learning of spatial data via multimodal embedding, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA. pp. 1993–2002. doi:[10.1145/3357384.3358001](https://doi.org/10.1145/3357384.3358001).
- [117] JEZDOVIÄ, I., NEDELJKOVIÄ, N., Å1/IVOJINOVIÄ, L., RADENKOVIÄ, B., Labus, A., 2018. Smart city: A system for measuring noise pollution. Smart Cities and Regional Development (SCRD) Journal 2, 79–85.
- [118] JezdoviÄ, I., PopoviÄ, S., RadenkoviÄ, M., Labus, A., BogdanoviÄ, Z., 2021. A crowdsensing platform for real-time monitoring and analysis of noise pollution in smart cities. Sustainable Computing: Informatics and Systems 31, 100588.
- [119] Ji, J., Wang, J., Huang, C., Wu, J., Xu, B., Wu, Z., Zhang, J., Zheng, Y., 2023a. Spatio-temporal self-supervised learning for traffic flow prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4356–4364.
- [120] Ji, J., Yu, F., Lei, M., 2023b. Self-supervised spatiotemporal graph neural networks with self-distillation for traffic prediction. IEEE Transactions on Intelligent Transportation Systems 24, 1580–1593. doi:[10.1109/TITS.2022.3219626](https://doi.org/10.1109/TITS.2022.3219626).
- [121] Ji, S., Zheng, Y., Li, T., 2016. Urban sensing based on human mobility, in: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1040–1051.
- [122] Ji, Z., Wang, H., Han, J., Pang, Y., 2020. Sman: Stacked multimodal attention network for cross-modal image–text retrieval. IEEE transactions on cybernetics 52, 1086–1097.
- [123] Jiang, R., Song, X., Huang, D., Song, X., Xia, T., Cai, Z., Wang, Z., Kim, K.S., Shibasaki, R., 2019. DeepUrbanEvent: A system for predicting citywide crowd dynamics at big events, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA. pp. 2114–2122. doi:[10.1145/3292500.3330654](https://doi.org/10.1145/3292500.3330654).
- [124] Jiang, Z., Chen, L., Zhou, B., Huang, J., Xie, T., Fan, X., Wang, C., 2021. ITV: Inferring traffic violation-prone locations with vehicle trajectories and road environment data. IEEE Systems Journal 15, 3913–3924. doi:[10.1109/JSYST.2020.3012743](https://doi.org/10.1109/JSYST.2020.3012743).
- [125] Jin, G., Liang, Y., Fang, Y., Shao, Z., Huang, J., Zhang, J., Zheng, Y., 2023a. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. IEEE Transactions on Knowledge and Data Engineering , 1–20doi:[10.1109/TKDE.2023.3333824](https://doi.org/10.1109/TKDE.2023.3333824).
- [126] Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J.Y., Shi, X., Chen, P.Y., Liang, Y., Li, Y.F., Pan, S., et al., 2023b. Time-llm: Time series forecasting by reprogramming large language models. arXiv preprint arXiv:2310.01728 .
- [127] Jin, M., Zhang, Y., Chen, W., Zhang, K., Liang, Y., Yang, B., Wang, J., Pan, S., Wen, Q., 2024. Position paper: What can large language models tell us about time series analysis, in: International Conference on Machine Learning (ICML 2024).
- [128] Johari, F., Peronato, G., Sadeghian, P., Zhao, X., Widén, J., 2020. Urban building energy modeling: State of the art and future prospects. Renewable and Sustainable Energy Reviews 128, 109902.
- [129] Kang, G.K., Gao, J.Z., Chiao, S., Lu, S., Xie, G., 2018. Air quality prediction: Big data and machine learning approaches. Int. J. Environ. Sci. Dev 9, 8–16.
- [130] Kang, Y., Zhang, F., Gao, S., Lin, H., Liu, Y., 2020. A review of urban physical environment sensing using street view imagery in public health studies. Annals of GIS 26, 261–275.
- [131] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D., 2020. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 .
- [132] Ke, J., Feng, S., Zhu, Z., Yang, H., Ye, J., 2021. Joint predictions of multi-modal ride-hailing demands: A deep multi-task multi-graph learning-based approach. Transportation Research Part C: Emerging Technologies 127, 103063.
- [133] Ke, S., Li, T., Song, L., Sun, Y., Sun, Q., Zhang, J., Zheng, Y., 2023. Spatio-temporal contrastive self-supervised learning for poi-level crowd flow inference. arXiv preprint arXiv:2309.03239 .
- [134] Keerthi Chandra, D., Wang, P., Leopold, J., Fu, Y., 2020. Collective embedding with feature importance: A unified approach for spatiotemporal network embedding, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Association for Computing Machinery, New York, NY, USA. pp. 615–624. doi:[10.1145/3340531.3412030](https://doi.org/10.1145/3340531.3412030).
- [135] Khan, I., Hou, F., Zakari, A., Tawiah, V., Ali, S.A., 2022. Energy use and urbanization as determinants of china's environmental quality: prospects of the paris climate agreement. Journal of Environmental Planning and Management 65, 2363–2386.
- [136] Khanna, S., Liu, P., Zhou, L., Meng, C., Rombach, R., Burke, M., Lobell, D., Ermon, S., 2023. Diffusionsat: A generative foundation model for satellite imagery. arXiv preprint arXiv:2312.03606 .
- [137] Kim, S.H., Lu, Y., Constantinou, G., Shahabi, C., Wang, G., Zimmermann, R., 2014. Mediaq: mobile multimedia management system, in: Proceedings of the 5th ACM Multimedia Systems Conference, pp. 224–235.
- [138] Kök, İ., Şimşek, M.U., Özdemir, S., 2017. A deep learning model for air quality prediction in smart cities, in: 2017 IEEE international conference on big data (big data), IEEE. pp. 1983–1990.
- [139] Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25.
- [140] Kruszyna, M., Śleszyński, P., Rychlewski, J., 2021. Dependencies between demographic urbanization and the agglomeration road traffic volumes: Evidence from poland. Land 10, 47.
- [141] Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S., 2023. Geochat: Grounded large vision-language model for remote sensing. arXiv preprint arXiv:2311.15826 .
- [142] Kwon, G., Cai, Z., Ravichandran, A., Bas, E., Bhotika, R., Soatto, S., 2022. Masked vision and language modeling for multi-modal representation learning. arXiv preprint arXiv:2208.02131 .
- [143] Lablack, M., Shen, Y., 2023. Spatio-temporal graph mixformer for traffic forecasting. Expert Systems with Applications 228, 120281.
- [144] Lai, S., Xu, Z., Zhang, W., Liu, H., Xiong, H., 2023a. Large language models as traffic signal control agents: Capacity and opportunity. arXiv preprint arXiv:2312.16044 .
- [145] Lai, S., Zhang, W., Liu, H., 2023b. A preference-aware meta-optimization framework for personalized vehicle energy consumption estimation, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA. pp. 4346–4356. doi:[10.1145/3580305.3599767](https://doi.org/10.1145/3580305.3599767).
- [146] Laud, T., Spokoyny, D., Corringham, T., Berg-Kirkpatrick, T., 2023. Climabench: A benchmark dataset for climate change text understanding in english. arXiv preprint arXiv:2301.04253 .
- [147] Le-Khac, P.H., Healy, G., Smeaton, A.F., 2020. Contrastive representation learning: A framework and review. Ieee Access 8, 193907–193934.
- [148] Li, F., Feng, J., Yan, H., Jin, G., Yang, F., Sun, F., Jin, D., Li, Y., 2023a. Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution. ACM Trans. Knowl. Discov. Data 17. URL: <https://doi.org/10.1145/3532611>, doi:[10.1145/3532611](https://doi.org/10.1145/3532611).
- [149] Li, H., Wang, X., Zhang, Z., Zhu, W., 2022a. Ood-gnn: Out-of-distribution generalized graph neural network. IEEE Transactions on Knowledge and Data Engineering .
- [150] Li, J., Cheong, T.S., Shen, J., Fu, D., 2019. Urbanization and rural–urban consumption disparity: Evidence from china. The Singapore Economic Review 64, 983–996.
- [151] Li, J., Li, D., Savarese, S., Hoi, S., 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 .
- [152] Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H., 2021a. Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems 34, 9694–9705.
- [153] Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., Liu, X., He, B., 2021b.

- A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*.
- [154] Li, T., Xin, S., Xi, Y., Tarkoma, S., Hui, P., Li, Y., 2022b. Predicting multi-level socioeconomic indicators from structural urban imagery, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 3282–3291.
- [155] Li, Y., Fan, Z., Yin, D., Jiang, R., Deng, J., Song, X., 2023c. HMGL: Heterogeneous multigraph contrastive learning for LBSN friend recommendation. *World Wide Web* 26, 1625–1648. doi:[10.1007/s11280-022-01092-5](https://doi.org/10.1007/s11280-022-01092-5).
- [156] Li, Y., Huang, W., Cong, G., Wang, H., Wang, Z., 2023d. Urban region representation learning with OpenStreetMap building footprints, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, pp. 1363–1373. doi:[10.1145/3580305.3599538](https://doi.org/10.1145/3580305.3599538).
- [157] Li, Y., Su, H., Demiryurek, U., Zheng, B., He, T., Shahabi, C., 2017a. Pare: A system for personalized route guidance, in: Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, p. 637–646. URL: <https://doi.org/10.1145/3038912.3052717>, doi:[10.1145/3038912.3052717](https://doi.org/10.1145/3038912.3052717).
- [158] Li, Y., Yu, R., Shahabi, C., Liu, Y., 2017b. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. arXiv preprint arXiv:1707.01926.
- [159] Li, Y., Yu, Y., Zhang, Q., Liang, C., He, P., Chen, W., Zhao, T., 2023e. Losparse: Structured compression of large language models based on low-rank and sparse approximation. arXiv preprint arXiv:2306.11222.
- [160] Li, Z., Xia, L., Tang, J., Xu, Y., Shi, L., Xia, L., Yin, D., Huang, C., 2024. Urbangpt: Spatio-temporal large language models. arXiv preprint arXiv:2403.00813.
- [161] Liang, L., Wang, Z., Li, J., 2019a. The effect of urbanization on environmental pollution in rapidly developing urban agglomerations. *Journal of cleaner production* 237, 117649.
- [162] Liang, Y., Ke, S., Zhang, J., Yi, X., Zheng, Y., 2018. Geoman: Multi-level attention networks for geo-sensory time series prediction., in: IJCAI, pp. 3428–3434.
- [163] Liang, Y., Ouyang, K., Jing, L., Ruan, S., Liu, Y., Zhang, J., Rosenblum, D.S., Zheng, Y., 2019b. Urbanfm: Inferring fine-grained urban flows, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 3132–3142.
- [164] Liang, Y., Ouyang, K., Sun, J., Wang, Y., Zhang, J., Zheng, Y., Rosenblum, D., Zimmermann, R., 2021. Fine-grained urban flow prediction, in: Proceedings of the Web Conference 2021, Association for Computing Machinery, New York, NY, USA, pp. 1833–1845. doi:[10.1145/3442381.3449792](https://doi.org/10.1145/3442381.3449792).
- [165] Liang, Y., Wen, H., Nie, Y., Jiang, Y., Jin, M., Song, D., Pan, S., Wen, Q., 2024a. Foundation models for time series analysis: A tutorial and survey. arXiv preprint arXiv:2403.14735.
- [166] Liang, Y., Xia, Y., Ke, S., Wang, Y., Wen, Q., Zhang, J., Zheng, Y., Zimmermann, R., 2023. Airformer: Predicting nationwide air quality in china with transformers, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 14329–14337.
- [167] Liang, Y., Zhao, Z., Ding, F., Tang, Y., He, Z., 2024b. Time-dependent trip generation for bike sharing planning: A multi-task memory-augmented graph neural network. *Information Fusion*, 102294.
- [168] Lin, X., Chen, Y., Li, G., Yu, Y., 2022. A causal inference look at unsupervised video anomaly detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1620–1629.
- [169] Lin, Y.T., Chen, Y.N., 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. arXiv preprint arXiv:2305.13711.
- [170] Lin, Z., Feng, J., Lu, Z., Li, Y., Jin, D., 2019. Deepstn+: Context-aware spatial-temporal neural network for crowd flow prediction in metropolis, in: Proceedings of the AAAI conference on artificial intelligence, pp. 1020–1027.
- [171] Liu, H., Dong, Z., Jiang, R., Deng, J., Deng, J., Chen, Q., Song, X., 2023a. Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, pp. 4125–4129.
- [172] Liu, H., Guo, Q., Zhu, H., Fu, Y., Zhuang, F., Ma, X., Xiong, H., 2023b. Characterizing and forecasting urban vibrancy evolution: A multi-view graph mining perspective. *ACM Transactions on Knowledge Discovery from Data* 17, 68:1–68:24. doi:[10.1145/3568683](https://doi.org/10.1145/3568683).
- [173] Liu, H., Han, J., Fu, Y., Li, Y., Chen, K., Xiong, H., 2022a. Unified route representation learning for multi-modal transportation recommendation with spatiotemporal pre-training. *The VLDB Journal — The International Journal on Very Large Data Bases* 32, 325–342. doi:[10.1007/s00778-022-00748-y](https://doi.org/10.1007/s00778-022-00748-y).
- [174] Liu, H., Li, C., Li, Y., Lee, Y.J., 2023c. Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744.
- [175] Liu, H., Li, T., Hu, R., Fu, Y., Gu, J., Xiong, H., 2019a. Joint representation learning for multi-modal transportation recommendation. Proceedings of the AAAI Conference on Artificial Intelligence 33, 1036–1043. doi:[10.1609/aaai.v33i01.33011036](https://doi.org/10.1609/aaai.v33i01.33011036).
- [176] Liu, H., Li, T., Hu, R., Fu, Y., Gu, J., Xiong, H., 2019b. Joint representation learning for multi-modal transportation recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1036–1043.
- [177] Liu, J., Li, T., Xie, P., Du, S., Teng, F., Yang, X., 2020a. Urban big data fusion based on deep learning: An overview. *Information Fusion* 53, 123–133.
- [178] Liu, J., Li, T., Yuan, Z., Huang, W., Xie, P., Huang, Q., 2022b. Symbolic aggregate approximation based data fusion model for dangerous driving behavior detection. *Information Sciences* 609, 626–643.
- [179] Liu, P., 2023. A review on remote sensing data fusion with generative adversarial networks (gan). *Authorea Preprints*.
- [180] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G., 2023d. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* 55, 1–35.
- [181] Liu, X., Hu, J., Li, Y., Diao, S., Liang, Y., Hooi, B., Zimmermann, R., 2023e. Unitime: A language-empowered unified model for cross-domain time series forecasting. arXiv preprint arXiv:2310.09751.
- [182] Liu, X., Xia, Y., Liang, Y., Hu, J., Wang, Y., Bai, L., Huang, C., Liu, Z., Hooi, B., Zimmermann, R., 2023f. Largest: A benchmark dataset for large-scale traffic forecasting. arXiv preprint arXiv:2306.08259.
- [183] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J., 2021. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering* 35, 857–876.
- [184] Liu, Y., Ao, X., Dong, L., Zhang, C., Wang, J., He, Q., 2022c. Spatiotemporal activity modeling via hierarchical cross-modal embedding. *IEEE Transactions on Knowledge and Data Engineering* 34, 462–474. doi:[10.1109/TKDE.2020.2983892](https://doi.org/10.1109/TKDE.2020.2983892).
- [185] Liu, Y., Ma, X., Shu, L., Yang, Q., Zhang, Y., Huo, Z., Zhou, Z., 2020b. Internet of things for noise mapping in smart cities: state of the art and future directions. *IEEE Network* 34, 112–118.
- [186] Liu, Y., Zhang, X., Ding, J., Xi, Y., Li, Y., 2023g. Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction, in: Proceedings of the ACM Web Conference 2023, pp. 4150–4160.
- [187] Lu, Y., To, H., Alfarrarjeh, A., Kim, S.H., Yin, Y., Zimmermann, R., Shahabi, C., 2016. Geouvg: User-generated mobile video dataset with fine granularity spatial metadata, in: Proceedings of the 7th international conference on multimedia systems, pp. 1–6.
- [188] Luo, H., Bao, Z., Cong, G., Culpepper, J.S., Khoa, N.L.D., 2021a. Let trajectories speak out the traffic bottlenecks. doi:[10.48550/arXiv.2107.12948](https://doi.org/10.48550/arXiv.2107.12948), arXiv:2107.12948.
- [189] Luo, W., Liu, Q., Zhou, Y., Ran, Y., Liu, Z., Hou, W., Pei, S., Lai, S., 2023. Spatiotemporal variations of “triple-demic” outbreaks of respiratory infections in the united states in the post-covid-19 era. *BMC Public Health* 23. URL: <https://doi.org/10.1186/s12889-023-17406-9>, doi:[10.1186/s12889-023-17406-9](https://doi.org/10.1186/s12889-023-17406-9).
- [190] Luo, Y., Ji, J., Sun, X., Cao, L., Wu, Y., Huang, F., Lin, C.W., Ji, R., 2021b. Dual-level collaborative transformer for image captioning, in: Proceedings of the AAAI conference on artificial intelligence, pp. 2286–2293.
- [191] Lv, Z., Xu, J., Zheng, K., Yin, H., Zhao, P., Zhou, X., 2018. Lc-rnn: A deep learning model for traffic speed prediction., in: IJCAI, p. 27th.
- [192] Ma, M., Xie, P., Teng, F., Wang, B., Ji, S., Zhang, J., Li, T., 2023a. HiST-GNN: Hierarchical spatio-temporal graph neural network for weather forecasting. *Information Sciences* 648, 119580.
- [193] Ma, Z., Meng, C., Ren, H., Ruan, S., Bao, J., Wang, X., Li, T., Zheng, Y.,

- 2023b. Sainf: Stay area inference of vehicles using surveillance camera records .
- [194] Mahmud, M., Kaiser, M.S., McGinnity, T.M., Hussain, A., 2021. Deep learning in mining biological data. *Cognitive computation* 13, 1–33.
- [195] Manvi, R., Khanna, S., Mai, G., Burke, M., Lobell, D., Ermon, S., 2023. Geollm: Extracting geospatial knowledge from large language models. arXiv preprint arXiv:2310.06213 .
- [196] Mao, Z., Li, Z., Li, D., Bai, L., Zhao, R., 2022. Jointly contrastive representation learning on road network and trajectory, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 1501–1510.
- [197] MediaQ Project, . MediaQ Project. URL: <http://mediaq1.cloudapp.net/home/>.
- [198] Miller, H.J., 2004. Tobler's first law and spatial analysis. *Annals of the association of American geographers* 94, 284–289.
- [199] Miller, P., de Barros, A.G., Kattan, L., Wirasinghe, S., 2016. Public transportation and sustainability: A review. *KSCE Journal of Civil Engineering* 20, 1076–1083.
- [200] Miyazawa, S., Song, X., Xia, T., Shibusaki, R., Kaneda, H., 2019. Integrating gps trajectory and topics from twitter stream for human mobility estimation. *Frontiers of Computer Science* 13, 460–470.
- [201] Mozaffari, S., Al-Jarrah, O.Y., Dianati, M., Jennings, P., Mouzakitis, A., 2020. Deep learning-based vehicle behavior prediction for autonomous driving applications: A review. *IEEE Transactions on Intelligent Transportation Systems* 23, 33–47.
- [202] Nakada, R., Gulluk, H.I., Deng, Z., Ji, W., Zou, J., Zhang, L., 2023. Understanding multimodal contrastive learning and incorporating unpaired data, in: International Conference on Artificial Intelligence and Statistics, PMLR. pp. 4348–4380.
- [203] Nam, K.W., Yang, K., 2022. Realroi: Discovering real regions of interest from geotagged photos. *IEEE Access* 10, 83489–83497.
- [204] Narayanan, M., Cherukuri, A.K., 2016. A study and analysis of recommendation systems for location-based social network (LBSN) with big data. *IIMB Management Review* 28, 25–30. doi:[10.1016/j.iimb.2016.01.001](https://doi.org/10.1016/j.iimb.2016.01.001).
- [205] Negash, N.M., Yang, J., 2023. Driver behavior modeling towards autonomous vehicles: Comprehensive review. *IEEE Access* .
- [206] Neuberg, L.G., 2003. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory* 19, 675–685.
- [207] Neupane, B., Horanont, T., Aryal, J., 2021. Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. *Remote Sensing* 13, 808.
- [208] Nourmohammadi, Z., Lilasathapornkit, T., Ashfaq, M., Gu, Z., Saberi, M., 2021. Mapping urban environmental performance with emerging data sources: A case of urban greenery and traffic noise in sydney, australia. *Sustainability* 13, 605.
- [209] Oh, G., Leblanc, D., Peng, H., 2019. Vehicle Energy Dataset (VED), a large-scale dataset for vehicle energy consumption research. *IEEE Transactions on Intelligent Transportation Systems* 23, 3302–3312. URL: <https://api.semanticscholar.org/CorpusID:146120975>.
- [210] Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 .
- [211] OpenAI, 2023. Gpt-4 technical report. [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
- [212] Quallane, A.A., Bakali, A., Bahnasse, A., Broumi, S., Talea, M., 2022. Fusion of engineering insights and emerging trends: Intelligent urban traffic management system. *Information Fusion* 88, 218–248.
- [213] Ouyang, K., Liang, Y., Liu, Y., Tong, Z., Ruan, S., Zheng, Y., Rosenblum, D.S., 2020. Fine-grained urban flow inference. *IEEE transactions on knowledge and data engineering* 34, 2755–2770.
- [214] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al., 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35, 27730–27744.
- [215] Ozbayoglu, A.M., Gudelek, M.U., Sezer, O.B., 2020. Deep learning for financial applications: A survey. *Applied Soft Computing* 93, 106384.
- [216] Pan, L., Ren, Q., Li, J., 2023. Spatial-temporal graph contrastive learning for urban traffic flow forecasting. *Authorea Preprints* .
- [217] Pan, Z., Liang, Y., Wang, W., Yu, Y., Zheng, Y., Zhang, J., 2019. Urban traffic prediction from spatio-temporal data using deep meta learning, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 1720–1730.
- [218] Pan, Z., Zhang, W., Liang, Y., Zhang, W., Yu, Y., Zhang, J., Zheng, Y., 2020. Spatio-temporal meta learning for urban traffic prediction. *IEEE Transactions on Knowledge and Data Engineering* 34, 1462–1476.
- [219] Park, K.G., Han, S., 2018. How use of location-based social network (LBSN) services contributes to accumulation of social capital. *Social Indicators Research* 136, 379–396. doi:[10.1007/s11205-016-1525-9](https://doi.org/10.1007/s11205-016-1525-9).
- [220] Pearl, J., Glymour, M., Jewell, N.P., 2016. Causal inference in statistics: A primer. 2016. Internet resource .
- [221] Peng, L., Zhang, Y., Shang, J., 2023. Generating efficient training data via ldm-based attribute manipulation. arXiv preprint arXiv:2307.07099 .
- [222] Perera, A., Nik, V.M., Chen, D., Scartezzini, J.L., Hong, T., 2020. Quantifying the impacts of climate change and extreme climate events on energy systems. *Nature Energy* 5, 150–159.
- [223] Piccialli, F., Canzaniello, M., Chiaro, D., Izzo, S., Qi, P., 2024. Graphite—generative reasoning and analysis for predictive handling in traffic efficiency. *Information Fusion* 106, 102265.
- [224] Psyllidis, A., Gao, S., Hu, Y., Kim, E.K., McKenzie, G., Purves, R., Yuan, M., Andris, C., 2022. Points of interest (poi): a commentary on the state of the art, challenges, and prospects for the future. *Computational Urban Science* 2, 20.
- [225] Qiang, Y., Wen, H., Wu, L., Mao, X., Wu, F., Wan, H., Hu, H., 2023. Modeling intra-and inter-community information for route and time prediction in last-mile delivery, in: 2023 IEEE 39th International Conference on Data Engineering (ICDE), IEEE. pp. 3106–3112.
- [226] Qu, H., Gong, Y., Chen, M., Zhang, J., Zheng, Y., Yin, Y., 2022. Forecasting fine-grained urban flows via spatio-temporal contrastive self-supervision. *IEEE Transactions on Knowledge and Data Engineering* .
- [227] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR. pp. 8748–8763.
- [228] Rajeh, T.M., Li, T., Li, C., Javed, M.H., Luo, Z., Alhaek, F., 2023. Modeling multi-regional temporal correlation with gated recurrent unit and multiple linear regression for urban traffic flow prediction. *Knowledge-Based Systems* 262. doi:[10.1016/j.knosys.2022.110237](https://doi.org/10.1016/j.knosys.2022.110237).
- [229] Rasenberg, M., Özürek, A., Dingemanse, M., 2020. Alignment in multimodal interaction: An integrative framework. *Cognitive science* 44, e12911.
- [230] Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T., 2023. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4088–4099.
- [231] Risal, A., Parajuli, P.B., Dash, P., Ouyang, Y., Linhoss, A., 2020. Sensitivity of hydrology and water quality to variation in land use and land cover data. *Agricultural Water Management* 241, 106366.
- [232] Roberts, J., Lüddecke, T., Das, S., Han, K., Albanie, S., 2023. Gpt4geo: How a language model sees the world's geography. arXiv preprint arXiv:2306.00020 .
- [233] Rogers, G., Koper, P., Ruktanonchai, C., , Ruktanonchai, N., Utazi, E., Woods, D., Cunningham, A., Tatem, A.J., Steele, J., Lai, S., Sorichetta, A., 2023. Exploring the relationship between temporal fluctuations in satellite nightlight imagery and human mobility across africa. *Remote Sensing* 15. URL: <https://doi.org/10.3390/rs15174252>, doi:[10.3390/rs15174252](https://doi.org/10.3390/rs15174252).
- [234] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695.
- [235] Ruan, S., Long, C., Ma, Z., Bao, J., He, T., Li, R., Chen, Y., Wu, S., Zheng, Y., 2022. Service time prediction for delivery tasks via spatial meta-learning, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3829–3837.
- [236] Schimanski, T., Bingler, J., Hyslop, C., Kraus, M., Leipold, M., 2023. Climatebert-netzero: Detecting and assessing net zero and reduction targets. arXiv preprint arXiv:2310.08096 .
- [237] Schrank, D., Eisele, B., Lomax, T., et al., 2019. Urban mobility report 2019 .

- [238] Shahid, N., Rappon, T., Berta, W., 2019. Applications of artificial neural networks in health care organizational decision-making: A scoping review. *PLoS one* 14, e0212356.
- [239] Shao, R., Yang, C., Li, Q., Zhu, Q., Zhang, Y., Li, Y., Liu, Y., Tang, Y., Liu, D., Yang, S., et al., 2023. Allspark: a multimodal spatiotemporal general model. *arXiv preprint arXiv:2401.00546*.
- [240] Shen, D., Zhang, L., Cao, J., Wang, S., 2018. Forecasting citywide traffic congestion based on social media. *Wireless Personal Communications* 103, 1037–1057.
- [241] Shen, S., Yao, Z., Li, C., Darrell, T., Keutzer, K., He, Y., 2023a. Scaling vision-language models with sparse mixture of experts. *arXiv preprint arXiv:2303.07226*.
- [242] Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y., 2023b. Hugging-gpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.
- [243] Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C., 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems* 28.
- [244] Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al., 2023. Large language models encode clinical knowledge. *Nature* 620, 172–180.
- [245] Sinha, R., Olsson, L.E., Frostell, B., 2019. Sustainable personal transport modes in a life cycle perspective—public or private? *Sustainability* 11, 7092.
- [246] Song, J., Tong, X., Wang, L., Zhao, C., Prishchepov, A.V., 2019a. Monitoring finer-scale population density in urban functional zones: A remote sensing data fusion approach. *Landscape and urban planning* 190, 103580.
- [247] Song, K., Yang, G., Wang, Q., Xu, C., Liu, J., Liu, W., Shi, C., Wang, Y., Zhang, G., Yu, X., et al., 2019b. Deep learning prediction of incoming rainfalls: An operational service for the city of beijing china, in: *2019 International Conference on Data Mining Workshops (ICDMW)*, IEEE. pp. 180–185.
- [248] Song, X., Kanasugi, H., Shibasaki, R., 2016. Deeptransport: Prediction and simulation of human mobility and transportation mode at a citywide level, in: *Proceedings of the twenty-fifth international joint conference on artificial intelligence*, pp. 2618–2624.
- [249] Song, X., Shibasaki, R., Yuan, N.J., Xie, X., Li, T., Adachi, R., 2017. DeepMob: Learning deep knowledge of human emergency behavior and mobility from big and heterogeneous data. *ACM Transactions on Information Systems* 35, 41:1–41:19. doi:[10.1145/3057280](https://doi.org/10.1145/3057280).
- [250] Steurer, M., Bayr, C., 2020. Measuring urban sprawl using land use data. *Land Use Policy* 97, 104799.
- [251] Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., Cilar, L., 2020. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, e1379.
- [252] Stratton, S.J., 2021. Population research: convenience sampling strategies. *Prehospital and disaster Medicine* 36, 373–374.
- [253] Sun, J., Zhang, J., Li, Q., Yi, X., Liang, Y., Zheng, Y., 2020. Predicting citywide crowd flows in irregular regions using multi-view graph convolutional networks. *IEEE Transactions on Knowledge and Data Engineering* 34, 2348–2359.
- [254] Sun, Y., Li, Y., Borozan, S., Wang, G., Qiu, J., Strbac, G., 2023. Battery swapping dispatch for self-sustained highway energy system based on spatiotemporal deep-learning traffic flow prediction. *IEEE Transactions on Industry Applications*.
- [255] Svikhnushina, E., Pu, P., 2023. Approximating online human evaluation of social chatbots with prompting, in: *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 268–281.
- [256] Talpur, A., Zhang, Y., 2018. A study of tourist sequential activity pattern through location based social network (LBSN), in: *2018 International Conference on Orange Technologies (ICOT)*, pp. 1–8. doi:[10.1109/ICOT.2018.8705895](https://doi.org/10.1109/ICOT.2018.8705895).
- [257] Tang, B., Pan, Z., Yin, K., Khateeb, A., 2019. Recent advances of deep learning in bioinformatics and computational biology. *Frontiers in genetics* 10, 214.
- [258] Tang, J., Xia, L., Hu, J., Huang, C., 2023a. Spatio-temporal meta contrastive learning, in: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 2412–2421.
- [259] Tang, J., Yang, Y., Wei, W., Shi, L., Su, L., Cheng, S., Yin, D., Huang, C., 2023b. Graphgpt: Graph instruction tuning for large language models. *arXiv preprint arXiv:2310.13023*.
- [260] Tang, R., Zhao, J., Liu, Y., Huang, X., Zhang, Y., Zhou, D., Ding, A., Nielsen, C.P., Wang, H., 2022. Air quality and health co-benefits of china's carbon dioxide emissions peaking before 2030. *Nature communications* 13, 1008.
- [261] Tedjopurnomo, D.A., Li, X., Bao, Z., Cong, G., Choudhury, F., Qin, A.K., 2021. Similar trajectory search with spatio-temporal deep representation learning. *ACM Transactions on Intelligent Systems and Technology (TIST)* 12, 1–26.
- [262] Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W., 2023. Large language models in medicine. *Nature medicine* 29, 1930–1940.
- [263] Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K.H., Poland, D., Borth, D., Li, L.J., 2015. The new data and new challenges in multimedia research. *arXiv* 1. doi:[arXiv:1503.01817](https://arxiv.org/abs/1503.01817).
- [264] Tian, B., Cao, Y., Zhang, Y., Xing, C., 2022. Debiasing nlu models via causal intervention and counterfactual reasoning, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11376–11384.
- [265] Tran, K., Sakla, W., Krim, H., 2021. Generative information fusion, in: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE. pp. 3990–3994.
- [266] Trisos, C.H., Merow, C., Pigot, A.L., 2020. The projected timing of abrupt ecological disruption from climate change. *Nature* 580, 496–501.
- [267] Tu, W., Li, Q., Fang, Z., Shaw, S.I., Zhou, B., Chang, X., 2016. Optimizing the locations of electric taxi charging stations: A spatial-temporal demand coverage approach. *Transportation Research Part C: Emerging Technologies* 65, 172–189.
- [268] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems* 30.
- [269] Veitch, V., Sridhar, D., Blei, D., 2020. Adapting text embeddings for causal inference, in: *Conference on Uncertainty in Artificial Intelligence*, PMLR. pp. 919–928.
- [270] Vu, D.D., To, H., Shin, W.Y., Shahabi, C., 2016. GeoSocialBound: An efficient framework for estimating social POI boundaries using spatio-textual information, in: *Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data*, ACM, San Francisco California. pp. 1–6. doi:[10.1145/2948649.2948652](https://doi.org/10.1145/2948649.2948652).
- [271] Wallin, M.T., Culpepper, W.J., Campbell, J.D., Nelson, L.M., Langer-Gould, A., Marrie, R.A., Cutter, G.R., Kaye, W.E., Wagner, L., Tremlett, H., et al., 2019. The prevalence of ms in the united states: a population-based estimate using health claims data. *Neurology* 92, e1029–e1040.
- [272] Wang, B., Lin, Y., Guo, S., Wan, H., 2021a. GSNet: Learning spatial-temporal correlations from geographical and semantic aspects for traffic accident risk forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 4402–4409. doi:[10.1609/aaai.v35i5.16566](https://doi.org/10.1609/aaai.v35i5.16566).
- [273] Wang, D., Liu, K., Johnson, P., Sun, L., Du, B., Fu, Y., 2021b. Deep human-guided conditional variational generative modeling for automated urban planning, in: *2021 IEEE international conference on data mining (ICDM)*, IEEE. pp. 679–688.
- [274] Wang, D., Peng, J., Tao, X., Duan, Y., 2024a. Boosting urban prediction tasks with domain-sharing knowledge via meta-learning. *Information Fusion* , 102324.
- [275] Wang, D., Wu, L., Zhang, D., Zhou, J., Sun, L., Fu, Y., 2023a. Human-instructed deep hierarchical generative learning for automated urban planning, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4660–4667.
- [276] Wang, H., Xiang, X., Fan, Y., Xue, J.H., 2024b. Customizing 360-degree panoramas through text-to-image diffusion models, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4933–4943.
- [277] Wang, H., Yu, Q., Liu, Y., Jin, D., Li, Y., 2021c. Spatio-temporal urban knowledge graph enabled mobility prediction. doi:[10.48550/arXiv.2111.03465](https://doi.org/10.48550/arXiv.2111.03465).
- [278] Wang, H., Yu, Q., Liu, Y., Jin, D., Li, Y., 2021d. Spatio-temporal urban knowledge graph enabled mobility prediction. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 1–24.
- [279] Wang, J., Zhang, X., Guo, Z., Lu, H., 2017a. Developing an early-

- warning system for air quality prediction and assessment of cities in china. *Expert systems with applications* 84, 102–116.
- [280] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al., 2023b. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.
- [281] Wang, Q., Lin, J., Zhou, K., Fan, J., Kwan, M.P., 2020a. Does urbanization lead to less residential energy consumption? a comparative study of 136 countries. *Energy* 202, 117765.
- [282] Wang, S., Cao, J., Philip, S.Y., 2020b. Deep learning for spatio-temporal data mining: A survey. *IEEE Transactions on Knowledge and Data Engineering* 34, 3681–3700.
- [283] Wang, S., He, L., Stenneth, L., Philip, S.Y., Li, Z., Huang, Z., 2016a. Estimating urban traffic congestions with multi-sourced data, in: 2016 17th IEEE International conference on mobile data management (MDM), IEEE. pp. 82–91.
- [284] Wang, S., He, L., Stenneth, L., Yu, P.S., Li, Z., 2015. Citywide traffic congestion estimation with social media, in: Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems, pp. 1–10.
- [285] Wang, S., Li, F., Stenneth, L., Yu, P.S., 2016b. Enhancing traffic congestion estimation with social media by coupled hidden markov model, in: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19–23, 2016, Proceedings, Part II 16, Springer. pp. 247–264.
- [286] Wang, S., Li, Y., Zhang, J., Meng, Q., Meng, L., Gao, F., 2020c. Pm2.5-gnn: A domain knowledge enhanced graph neural network for pm2.5 forecasting, in: Proceedings of the 28th international conference on advances in geographic information systems, pp. 163–166.
- [287] Wang, S., Zhang, J., Li, J., Miao, H., Cao, J., 2021e. Traffic accident risk prediction via multi-view multi-task spatio-temporal networks. *IEEE Transactions on Knowledge and Data Engineering*.
- [288] Wang, S., Zhang, J., Li, J., Miao, H., Cao, J., 2023c. Traffic accident risk prediction via multi-view multi-task spatio-temporal networks. *IEEE Transactions on Knowledge and Data Engineering* 35, 12323–12336. doi:[10.1109/TKDE.2021.3135621](https://doi.org/10.1109/TKDE.2021.3135621).
- [289] Wang, S., Zhang, X., Cao, J., He, L., Stenneth, L., Yu, P.S., Li, Z., Huang, Z., 2017b. Computing urban traffic congestions by incorporating sparse gps probe data and social media data. *ACM Transactions on Information Systems (TOIS)* 35, 1–30.
- [290] Wang, T., Huang, J., Zhang, H., Sun, Q., 2020d. Visual commonsense r-cnn, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10760–10770.
- [291] Wang, X., Fang, M., Zeng, Z., Cheng, T., 2023d. Where would i go next? large language models as human mobility predictors. *arXiv preprint arXiv:2308.15197*.
- [292] Wang, Y., Zhu, D., 2024. A hypergraph-based hybrid graph convolutional network for intracity human activity intensity prediction and geographic relationship interpretation. *Information Fusion* 104, 102149.
- [293] Wang, Z., Peng, Z., Wang, S., Song, Q., 2022. Personalized long-distance fuel-efficient route recommendation through historical trajectories mining, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, ACM, Virtual Event AZ USA. pp. 1072–1080. doi:[10.1145/3488560.3498512](https://doi.org/10.1145/3488560.3498512).
- [294] Wang, Z., Wang, H., Qin, F., Han, Z., Miao, C., 2020e. Mapping an urban boundary based on multi-temporal sentinel-2 and POI data: A case study of zhengzhou city. *Remote Sensing* 12, 4103. doi:[10.3390/rs12244103](https://doi.org/10.3390/rs12244103).
- [295] Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y., 2021f. Simvilm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.
- [296] Wei, X., Zhang, T., Li, Y., Zhang, Y., Wu, F., 2020. Multi-modality cross attention network for image and sentence matching, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10941–10950.
- [297] Wen, H., Lin, Y., Hu, Y., Wu, F., Xia, M., Zhang, X., Wu, L., Hu, H., Wan, H., 2023a. Modeling spatial-temporal constraints and spatial-transfer patterns for couriers' package pick-up route prediction. *IEEE Transactions on Intelligent Transportation Systems*.
- [298] Wen, H., Lin, Y., Wu, F., Wan, H., Sun, Z., Cai, T., Liu, H., Guo, S., Zheng, J., Song, C., et al., 2023b. Enough waiting for the couriers: Learning to estimate package pick-up arrival time from couriers' spatial-temporal behaviors. *ACM Transactions on Intelligent Systems and Technology* 14, 1–22.
- [299] Wen, H., Lin, Y., Xia, Y., Wan, H., Wen, Q., Zimmermann, R., Liang, Y., 2023c. Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models, in: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, pp. 1–12.
- [300] Wen, Y., Bein, D., Phoha, S., 2014. Dynamic clustering of multi-modal sensor networks in urban scenarios. *Information Fusion* 15, 130–140.
- [301] Wu, F., Li, Z., Lee, W.C., Wang, H., Huang, Z., 2015. Semantic annotation of mobility data using social media, in: Proceedings of the 24th international conference on world wide web, pp. 1253–1263.
- [302] Wu, G., Ding, Y., Li, Y., Bao, J., Zheng, Y., Luo, J., 2017. Mining spatio-temporal reachable regions over massive trajectory data, in: 2017 IEEE 33rd International Conference on Data Engineering (ICDE), pp. 1283–1294. doi:[10.1109/ICDE.2017.171](https://doi.org/10.1109/ICDE.2017.171).
- [303] Wu, H., Hao, Y., Weng, J.H., 2019a. How does energy consumption affect china's urbanization? new evidence from dynamic threshold panel models. *Energy policy* 127, 24–38.
- [304] Wu, L., Liu, J., Lou, J., Hu, H., Zheng, J., Wen, H., Song, C., He, S., 2023a. G2ptl: A pre-trained model for delivery address and its applications in logistics system. *arXiv preprint arXiv:2304.01559*.
- [305] Wu, L., Wen, H., Hu, H., Mao, X., Xia, Y., Shan, E., Zhen, J., Lou, J., Liang, Y., Yang, L., et al., 2023b. Lade: The first comprehensive last-mile delivery dataset from industry. *arXiv preprint arXiv:2306.10675*.
- [306] Wu, S., Irsoy, O., Lu, S., Dabrowski, V., Dredze, M., Lehrmann, S., Kambadur, P., Rosenberg, D., Mann, G., 2023c. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
- [307] Wu, T., Xiang, L., Gong, J., 2016. Updating road networks by local renewal from gps trajectories. *ISPRS International Journal of Geoinformation* 5, 163.
- [308] Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C., 2019b. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*.
- [309] Xi, Y., Li, T., Wang, H., Li, Y., Tarkoma, S., Hui, P., 2022. Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests, in: Proceedings of the ACM Web Conference 2022, ACM, Virtual Event, Lyon France. pp. 3308–3316. doi:[10.1145/3485447.3512149](https://doi.org/10.1145/3485447.3512149).
- [310] Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al., 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.
- [311] Xia, L., Huang, C., Xu, Y., Dai, P., Bo, L., Zhang, X., Chen, T., 2021. Spatial-temporal sequential hypergraph network for crime prediction with dynamic multiplex relation learning, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization.
- [312] Xia, Y., Liang, Y., Wen, H., Liu, X., Wang, K., Zhou, Z., Zimmermann, R., 2024. Deciphering spatio-temporal graph forecasting: A causal lens and treatment. *Advances in Neural Information Processing Systems* 36.
- [313] Xiao, C., Zhou, J., Huang, J., Zhu, H., Xu, T., Dou, D., Xiong, H., 2023. A contextual master-slave framework on urban region graph for urban village detection, in: 2023 IEEE 39th International Conference on Data Engineering (ICDE), IEEE Computer Society. pp. 736–748. doi:[10.1109/ICDE55515.2023.00062](https://doi.org/10.1109/ICDE55515.2023.00062).
- [314] Xie, J., Liang, Y., Liu, J., Xiao, Y., Wu, B., Ni, S., 2023. Quert: Continual pre-training of language model for query understanding in travel domain search. *arXiv preprint arXiv:2306.06707*.
- [315] Xie, P., Li, T., Liu, J., Du, S., Yang, X., Zhang, J., 2020. Urban flow prediction from spatiotemporal data using machine learning: A survey. *Information Fusion* 59, 1–12.
- [316] Xu, D., Chen, Y., Cui, N., Li, J., 2023a. Towards multi-dimensional knowledge-aware approach for effective community detection in LBSN. *World Wide Web* 26, 1435–1458. doi:[10.1007/s11280-022-01101-7](https://doi.org/10.1007/s11280-022-01101-7).
- [317] Xu, F., Zhang, J., Gao, C., Feng, J., Li, Y., 2023b. Urban generative intelligence (ugi): A foundational platform for agents in embodied city environment. *arXiv preprint arXiv:2312.11813*.
- [318] Xu, J., Wang, S., Ying, N., Xiao, X., Zhang, J., Jin, Z., Cheng, Y., Zhang, G., 2023c. Dynamic graph neural network with adaptive edge attributes

- for air quality prediction: A case study in china. *Heliyon* 9.
- [319] Xu, M., Yoon, S., Fuentes, A., Park, D.S., 2023d. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition* , 109347.
- [320] Xu, X., Wei, Y., Wang, P., Luo, X., Zhou, F., Trajcevski, G., 2023e. Diffusion probabilistic modeling for fine-grained urban traffic flow inference with relaxed structural constraint, in: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE. pp. 1–5.
- [321] Xu, Z., Peng, J., Liu, Y., Qiu, S., Zhang, H., Dong, J., 2023f. Exploring the combined impact of ecosystem services and urbanization on sdgs realization. *Applied Geography* 153, 102907.
- [322] Xue, H., Voutharaja, B.P., Salim, F.D., 2022. Leveraging language foundation models for human mobility forecasting, in: *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pp. 1–9.
- [323] Yan, Y., Wen, H., Zhong, S., Chen, W., Chen, H., Wen, Q., Zimmermann, R., Liang, Y., 2023. When urban region profiling meets large language models. *arXiv preprint arXiv:2310.18340* .
- [324] Yang, J., Ye, X., Wu, B., Gu, Y., Wang, Z., Xia, D., Huang, J., 2022. DuARE: Automatic road extraction with aerial images and trajectory data at baidu maps, in: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, NY, USA. pp. 4321–4331. doi:[10.1145/3534678.3539029](https://doi.org/10.1145/3534678.3539029).
- [325] Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H., 2023a. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys* 56, 1–39.
- [326] Yang, M., Guo, T., Zhu, T., Tjuawinata, I., Zhao, J., Lam, K.Y., 2023b. Local differential privacy and its applications: A comprehensive survey. *Computer Standards & Interfaces* , 103827.
- [327] Yang, W., Ueda, A., Sugiura, K., 2023c. Multimodal encoder with gated cross-attention for text-vqa tasks, in: *29th Annual Conference of the Language Processing Society*, pp. 1580–1585.
- [328] Yao, D., Hu, H., Du, L., Cong, G., Han, S., Bi, J., 2022. Trajagt: A graph-based long-term dependency modeling approach for trajectory similarity computation, in: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2275–2285.
- [329] Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., Gong, P., Ye, J., Li, Z., 2018. Deep multi-view spatial-temporal network for taxi demand prediction, in: *Proceedings of the AAAI conference on artificial intelligence*.
- [330] Yao, Y., Jiang, L., 2021. Urbanization forces driving rural urban income disparity: Evidence from metropolitan areas in china. *Journal of Cleaner Production* 312, 127748.
- [331] Yao, Z., Wu, X., Li, C., Youn, S., He, Y., 2023. Zeroquant-v2: Exploring post-training quantization in llms from comprehensive study to low rank compensation. *arXiv preprint arXiv:2303.08302* .
- [332] Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al., 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178* .
- [333] Ye, Y., Xie, H., Fang, J., Jiang, H., Wang, D., 2019. Daily accessed street greenery and housing price: Measuring economic performance of human-scale streetscapes via new urban data. *Sustainability* 11, 1741.
- [334] Yi, X., Duan, Z., Li, R., Zhang, J., Li, T., Zheng, Y., 2020. Predicting fine-grained air quality based on deep neural networks. *IEEE Transactions on Big Data* 8, 1326–1339.
- [335] Yi, X., Zhang, J., Wang, Z., Li, T., Zheng, Y., 2018. Deep distributed fusion network for air quality prediction, in: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 965–973.
- [336] Yin, B., Xie, J., Qin, Y., Ding, Z., Feng, Z., Li, X., Lin, W., 2023a. Heterogeneous knowledge fusion: A novel approach for personalized recommendation via llm, in: *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 599–601.
- [337] Yin, Y., Hu, W., Tran, A., Zhang, Y., Wang, G., Kruppa, H., Zimmermann, R., Ng, S.K., 2023b. Multimodal deep learning for robust road attribute detection. *ACM Transactions on Spatial Algorithms and Systems* .
- [338] Yin, Y., Liu, Z., Zhang, Y., Wang, S., Shah, R.R., Zimmermann, R., 2019. Gps2vec: Towards generating worldwide gps embeddings, in: *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 416–419.
- [339] Yin, Y., Varadarajan, J., Wang, G., Wang, X., Sahrawat, D., Zimmermann, R., Ng, S.K., 2020. A multi-task learning framework for road attribute updating via joint analysis of map data and GPS traces, in: *Proceedings of The Web Conference 2020*, Association for Computing Machinery, New York, NY, USA. pp. 2662–2668. doi:[10.1145/3366423.3380021](https://doi.org/10.1145/3366423.3380021).
- [340] Yin, Y., Zhang, Y., Liu, Z., Liang, Y., Wang, S., Shah, R.R., Zimmermann, R., 2021. Learning multi-context aware location representations from large-scale geotagged images, in: *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 899–907.
- [341] Yonekura, K., Hattori, H., Suzuki, T., 2018. Short-term local weather forecast using dense weather station by deep neural network, in: *2018 IEEE international conference on big data (big data)*, IEEE. pp. 1683–1690.
- [342] You, J., Muhammad, A.S., He, X., Xie, T., Wang, Z., Fan, X., Yu, Z., Chen, L., Wang, C., 2022. Panda: predicting road risks after natural disasters leveraging heterogeneous urban data. *CCF Transactions on Pervasive Computing and Interaction* 4, 393–407.
- [343] Yu, B., Yin, H., Zhu, Z., 2017. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875* .
- [344] Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y., 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* .
- [345] Yu, X., Chen, Z., Ling, Y., Dong, S., Liu, Z., Lu, Y., 2023. Temporal data meets llm-explainable financial time series forecasting. *arXiv preprint arXiv:2306.11025* .
- [346] Yu, X., Shi, S., Xu, L., 2021. A spatial-temporal graph attention network approach for air temperature forecasting. *Applied Soft Computing* 113, 107888.
- [347] Yu, Z., Xu, H., Yang, Z., Guo, B., 2015. Personalized travel package with multi-point-of-interest recommendation based on crowdsourced user footprints. *IEEE Transactions on Human-Machine Systems* 46, 151–158.
- [348] Yuan, H., Li, G., 2021. A survey of traffic prediction: from spatio-temporal data to intelligent transportation. *Data Science and Engineering* 6, 63–85.
- [349] Yuan, H., Li, G., Bao, Z., Feng, L., 2021a. An effective joint prediction model for travel demands and traffic flows, in: *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, IEEE. pp. 348–359.
- [350] Yuan, J., Zheng, Y., Xie, X., Sun, G., 2011. Driving with knowledge from the physical world, in: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 316–324.
- [351] Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., Huang, Y., 2010. T-drive: driving directions based on taxi trajectories, in: *Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems*, pp. 99–108.
- [352] Yuan, Q., Zhang, W., Zhang, C., Geng, X., Cong, G., Han, J., 2017. Pred: Periodic region detection for mobility modeling of social media users, in: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 263–272.
- [353] Yuan, X., Shi, J., Gu, L., 2021b. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications* 169, 114417.
- [354] Yuan, Y., Ding, J., Wang, H., Jin, D., Li, Y., 2022. Activity trajectory generation via modeling spatiotemporal dynamics, in: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4752–4762.
- [355] Yuan, Y., Wang, H., Ding, J., Jin, D., Li, Y., 2023. Learning to simulate daily activities via modeling dynamic human needs, in: *Proceedings of the ACM Web Conference 2023*, pp. 906–916.
- [356] Yuanshao, Z., Ye, Y., Zhang, S., Zhao, X., James, J.Y., 2023. Diff traj: Generating gps trajectory with diffusion probabilistic model, in: *Proceedings of the 37th Annual Conference on Neural Information Processing Systems*.
- [357] Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., Gao, Y., 2021a. A survey on federated learning. *Knowledge-Based Systems* 216, 106775.

- [358] Zhang, C., Zhang, Y., Shao, Q., Li, B., Lv, Y., Piao, X., Yin, B., 2023a. Chattraffic: Text-to-traffic generation via diffusion model. arXiv preprint arXiv:2311.16203 .
- [359] Zhang, D., Zhang, H., Tang, J., Hua, X.S., Sun, Q., 2020a. Causal intervention for weakly-supervised semantic segmentation. Advances in Neural Information Processing Systems 33, 655–666.
- [360] Zhang, H., Dai, L., 2018. Mobility prediction: A survey on state-of-the-art schemes and future applications. IEEE access 7, 802–822.
- [361] Zhang, J., Xie, Y., Ding, W., Wang, Z., 2023b. Cross on cross attention: Deep fusion transformer for image captioning. IEEE Transactions on Circuits and Systems for Video Technology .
- [362] Zhang, J., Zheng, Y., Qi, D., 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction, in: Proceedings of the AAAI conference on artificial intelligence.
- [363] Zhang, J., Zheng, Y., Qi, D., Li, R., Yi, X., 2016. Dnn-based prediction model for spatio-temporal data, in: Proceedings of the 24th ACM SIGSPATIAL international conference on advances in geographic information systems, pp. 1–4.
- [364] Zhang, J., Zheng, Y., Qi, D., Li, R., Yi, X., Li, T., 2018. Predicting citywide crowd flows using deep spatio-temporal residual networks. Artificial Intelligence 259, 147–166.
- [365] Zhang, L., Geng, X., Qin, Z., Wang, H., Wang, X., Zhang, Y., Liang, J., Wu, G., Song, X., Wang, Y., 2022a. Multi-modal graph interaction for multi-graph convolution network in urban spatiotemporal forecasting. Sustainability 14, 12397.
- [366] Zhang, L., Long, C., Cong, G., 2023c. Region embedding with intra and inter-view contrastive learning. IEEE Transactions on Knowledge and Data Engineering 35, 9031–9036. doi:10.1109/TKDE.2022.3220874.
- [367] Zhang, L., Rao, A., Agrawala, M., 2023d. Adding conditional control to text-to-image diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847.
- [368] Zhang, M., Li, T., Li, Y., Hui, P., 2021b. Multi-view joint graph representation learning for urban region embedding, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Yokohama, Yokohama, Japan, pp. 4431–4437.
- [369] Zhang, P., Dong, X., Wang, B., Cao, Y., Xu, C., Ouyang, L., Zhao, Z., Ding, S., Zhang, S., Duan, H., Zhang, W., Yan, H., Zhang, X., Li, W., Li, J., Chen, K., He, C., Zhang, X., Qiao, Y., Lin, D., Wang, J., 2023e. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. arXiv:2309.15112 .
- [370] Zhang, Q., Han, Y., Li, V.O., Lam, J.C., 2022b. Deep-air: A hybrid cnn-lstm framework for fine-grained air pollution estimation and forecast in metropolitan cities. IEEE Access 10, 55818–55841.
- [371] Zhang, R., Kennard, N.N., Smith, D., McFarland, D., McCallum, A., Keith, K., 2023f. Causal matching with text embeddings: A case study in estimating the causal effects of peer review policies, in: Findings of the Association for Computational Linguistics: ACL 2023, pp. 1284–1297.
- [372] Zhang, S., Fatih, D., Abdulqadir, F., Schwarz, T., Ma, X., 2022c. Extended vehicle energy dataset (eved): an enhanced large-scale dataset for deep learning on vehicle trip energy consumption. arXiv preprint arXiv:2203.08630 .
- [373] Zhang, X., Gong, Y., Zhang, C., Wu, X., Guo, Y., Lu, W., Zhao, L., Dong, X., 2023g. Spatio-temporal fusion and contrastive learning for urban flow prediction. Knowledge-Based Systems 282, 111104.
- [374] Zhang, X., Han, L., Wei, H., Tan, X., Zhou, W., Li, W., Qian, Y., 2022d. Linking urbanization and air quality together: A review and a perspective on the future sustainable urban development. Journal of Cleaner Production 346, 130988.
- [375] Zhang, X., Huang, C., Xu, Y., Xia, L., Dai, P., Bo, L., Zhang, J., Zheng, Y., 2021c. Traffic flow forecasting with spatial-temporal graph diffusion network, in: Proceedings of the AAAI conference on artificial intelligence, pp. 15008–15015.
- [376] Zhang, X., Liu, L., Lan, M., Song, G., Xiao, L., Chen, J., 2022e. Interpretable machine learning models for crime prediction. Computers, Environment and Urban Systems 94, 101789.
- [377] Zhang, Y., Li, Q., Tu, W., Mai, K., Yao, Y., Chen, Y., 2019. Functional urban land use recognition integrating multi-source geospatial data and cross-correlations. Computers, Environment and Urban Systems 78, 101374.
- [378] Zhang, Y., Yin, Y., Zimmermann, R., Wang, G., Varadarajan, J., Ng, S.K., 2020b. An enhanced gan model for automatic satellite-to-map image conversion. IEEE Access 8, 176704–176716.
- [379] Zhao, B., Zhang, S., Xu, C., Sun, Y., Deng, C., 2021a. Deep fake geography? when geospatial data encounter artificial intelligence. Cartography and Geographic Information Science 48, 338–352.
- [380] Zhao, K., Cong, G., Li, X., 2020. Pgeotopic: A distributed solution for mining geographical topic models. IEEE Transactions on Knowledge and Data Engineering 34, 881–893.
- [381] Zhao, K., Cong, G., Sun, A., 2016a. Annotating points of interest with geo-tagged tweets, in: Proceedings of the 25th ACM International Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA, pp. 417–426. doi:10.1145/2983323.2983850.
- [382] Zhao, K., Liu, Y., Hao, S., Lu, S., Liu, H., Zhou, L., 2021b. Bounding boxes are all we need: street view image classification via context encoding of detected buildings. IEEE Transactions on Geoscience and Remote Sensing 60, 1–17.
- [383] Zhao, K., Liu, Y., Yuan, Q., Chen, L., Chen, Z., Cong, G., 2016b. Towards personalized maps: Mining user preferences from geo-textual data. Proceedings of the VLDB Endowment 9, 1545–1548.
- [384] Zhao, L., Gao, Y., Ye, J., Chen, F., Ye, Y., Lu, C.T., Ramakrishnan, N., 2022. Spatio-temporal event forecasting using incremental multi-source feature learning. ACM Transactions on Knowledge Discovery from Data 16, 1–28. doi:10.1145/3464976.
- [385] Zhao, L., Sun, Q., Ye, J., Chen, F., Lu, C.T., Ramakrishnan, N., 2015. Multi-task learning for spatio-temporal event forecasting, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, p. 1503–1512. URL: <https://doi.org/10.1145/2783258.2783377>, doi:10.1145/2783258.2783377.
- [386] Zhao, P., Xu, X., Liu, Y., Sheng, V.S., Zheng, K., Xiong, H., 2017. Photo2trip: Exploiting visual contents in geo-tagged photos for personalized tour recommendation, in: Proceedings of the 25th ACM international conference on Multimedia, pp. 916–924.
- [387] Zhao, S., Yu, Y., 2017. Effect of short-term regional traffic restriction on urban submicron particulate pollution. Journal of Environmental Sciences 55, 86–99.
- [388] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al., 2023. A survey of large language models. arXiv preprint arXiv:2303.18223 .
- [389] Zheng, S., Trott, A., Srinivasa, S., Parkes, D.C., Socher, R., 2022. The ai economist: Taxation policy design via two-level deep multiagent reinforcement learning. Science advances 8, eabk2607.
- [390] Zheng, Y., 2015. Methodologies for cross-domain data fusion: An overview. IEEE transactions on big data 1, 16–34.
- [391] Zheng, Y., Capra, L., Wolfson, O., Yang, H., 2014a. Urban computing: concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology (TIST) 5, 1–55.
- [392] Zheng, Y., Chen, X., Jin, Q., Chen, Y., Qu, X., Liu, X., Chang, E., Ma, W.Y., Rui, Y., Sun, W., 2014b. A cloud-based knowledge discovery system for monitoring fine-grained air quality. MSR-TR-2014-40, Tech. Rep. .
- [393] Zheng, Y., Gao, C., Li, X., He, X., Li, Y., Jin, D., 2021. Disentangling user interest and conformity for recommendation with causal embedding, in: Proceedings of the Web Conference 2021, pp. 2980–2991.
- [394] Zheng, Y., Li, Q., Chen, Y., Xie, X., Ma, W.Y., 2008. Understanding mobility based on gps data, in: Proceedings of the 10th international conference on Ubiquitous computing, pp. 312–321.
- [395] Zheng, Y., Lin, Y., Zhao, L., Wu, T., Jin, D., Li, Y., 2023a. Spatial planning of urban communities via deep reinforcement learning. Nature Computational Science 3, 748–762.
- [396] Zheng, Y., Liu, F., Hsieh, H.P., 2013. U-air: When urban air quality inference meets big data, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1436–1444.
- [397] Zheng, Y., Su, H., Ding, J., Jin, D., Li, Y., 2023b. Road planning for slums via deep reinforcement learning, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, p. 5695–5706.
- [398] Zheng, Y., Xie, X., Ma, W.Y., et al., 2010. Geolife: A collaborative social networking service among user, location and trajectory .

- [399] Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., Li, T., 2015. Forecasting fine-grained air quality based on big data, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, pp. 2267–2276. doi:[10.1145/2783258.2788573](https://doi.org/10.1145/2783258.2788573).
- [400] Zheng, Y., Zhang, L., Xie, X., Ma, W.Y., 2009. Mining interesting locations and travel sequences from gps trajectories, in: Proceedings of the 18th international conference on World wide web, pp. 791–800.
- [401] Zheng, Y., Zhong, L., Wang, S., Yang, Y., Gu, W., Zhang, J., Wang, J., 2023c. Diffuflow: Robust fine-grained urban flow inference with denoising diffusion model, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, pp. 3505–3513.
- [402] Zhou, T., Niu, P., Wang, X., Sun, L., Jin, R., 2023a. One fits all: Power general time series analysis by pretrained lm. arXiv preprint arXiv:2302.11939 .
- [403] Zhou, Z., Huang, Q., Yang, K., Wang, K., Wang, X., Zhang, Y., Liang, Y., Wang, Y., 2023b. Maintaining the status quo: Capturing invariant relations for ood spatiotemporal learning .
- [404] Zhou, Z., Lin, Y., Jin, D., Li, Y., 2024. Large language model for participatory urban planning. arXiv preprint arXiv:2402.17161 .
- [405] Zhu, S., Wang, D., Liu, L., Wang, Y., Guo, D., 2020. Inferring region significance by using multi-source spatial data. Neural Computing and Applications 32, 6523–6531.
- [406] Zohourianshahzadi, Z., Kalita, J.K., 2022. Neural attention for image captioning: review of outstanding methods. Artificial Intelligence Review 55, 3833–3862.
- [407] Zoph, B., Cubuk, E.D., Ghiasi, G., Lin, T.Y., Shlens, J., Le, Q.V., 2020. Learning data augmentation strategies for object detection, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16, Springer, pp. 566–583.
- [408] Zou, X., Huang, J., Hao, X., Yang, Y., Wen, H., Yan, Y., Huang, C., Liang, Y., 2024. Learning geospatial region embedding with heterogeneous graph. arXiv preprint arXiv:2405.14135 .