[DM-03-074] Modeling Semi-Structured and Unstructured Spatial Data

Abstract

This chapter surveys semi-structured and unstructured geospatial data, emphasizing their formats, challenges, and analytical approaches. Semi-structured data formats, such as JSON, do not follow rigid schemas but retain internal organization that supports spatial processing. These formats underpin many widely used datasets, including OpenStreetMap, and can represent both object-based and network-based spatial models. Unstructured data, including text, imagery, sensor streams, and point clouds, lack standardized formatting and must be transformed or enriched before spatial analysis is possible. For instance, crowdsourced or drone-collected imagery can be processed using Structure from Motion (SfM) to reconstruct 3D surfaces and terrain models. Textual data, such as social media posts or institutional reports, can be mined for geographic content using natural language processing techniques like named entity recognition and geoparsing. The chapter also considers recent developments in AI, including deep learning methods for image classification, segmentation of point clouds, and modeling spatiotemporal patterns from sensor data. Finally, it discusses the emerging role of multimodal models that integrate visual and textual information in geospatial workflows. Together, these tools and methods enable the use of increasingly diverse data sources in spatial analysis, broadening both the scope and depth of geographic inquiry.

Keywords: natural language processing, semi-structured text, unstructured text

Author & citation

Adams, B. (2025). The Geographic Information Science & Technology Body of Knowledge (Issue 2, 2025 Edition), John P. Wilson (ed.). DOI: 10.22224/gistbok/2025.2.11.

Explanation

1. Introduction

Students and practitioners of GIS will be familiar with structured spatial data in the form of vector, raster, and network data types, which are designed to be read by and analyzed in GIS applications. For a vector data set there is a fixed relational database schema with well-defined attributes and a spatial reference system. Likewise, raster data have well-defined coordinate systems and attributes for pixel values. Beyond that, however, there is a wealth of data that can be useful for geographical analysis that comes in various semi-structured or unstructured data formats.

Semi-structured data is spatial data that does not conform to rigid schema. Examples of semi-structured geospatial data include GPS traces, user-generated OpenStreetMap data, and earthquake and weather event data (Mooney and Minghini 2017; Mueller 2019). Unstructured spatial data comprises an even broader category that takes many forms from raw text and images to sensor data. We use unstructured data in geospatial analysis when it has some spatial metadata associated with it or it otherwise can be linked to known geographic entities. In some cases the inherent level of spatial detail associated with unstructured data is less than what you might find in more structured spatial data, but the richness of the attributes or the timeliness of the data give it added value.

Semi-structured and unstructured spatial data do not conform neatly to traditional vector or raster models, which are grounded in fixed schema representations. Instead, interpreting such data frequently requires the application of higher-level conceptual frameworks, such as field-based, object-based, or network models, that abstract spatial phenomena independently of storage format. These conceptual models provide the theoretical foundation for structuring meaning from heterogeneous data sources, including imagery, text, point clouds, and sensor streams.

In this entry, we discuss examples of semi-structured and unstructured geospatial data, looking at how the data are analyzed and some of their key applications. We also discuss the increasing role of artificial intelligence (AI) in managing and extracting meaning from these kinds of data.

2. Semi-Structured Geospatial Data

The key feature that distinguishes semi-structured data from traditional structured GIS data is the lack of a rigid relational schema. This means that new fields and attribute values can be added arbitrarily, and even the data types of the individual fields need not be fixed. The structured part of semi-structured, comes from the fact that the data is formatted using a well-known web data interchange format such as CSV (comma separated values), JSON (JavaScript Object Notation), XML (extensible markup language), or GPX (GPS exchange format). Furthermore, while the spatial attributes of semi-structured data sets are usually specified in longitude and latitude (and can be assumed to be WGS84), they often do not include explicit projection information.

The popularity of web mapping has led a rise in the use of GeoJSON, a standard form of JSON designed for interchanging point, line, and polygon vector data. GeoJSON feature geometry is defined by a set of coordinates with longitude and latitude units in decimal degrees using the WGS 84 datum. All modern web mapping toolkits can render styled GeoJSON layers over a base map. As a result, many open GIS data sets are now shared in GeoJSON format. Because any GeoJSON file is also JSON, users can add custom attribute fields to geographic features in a schema-less manner. Furthermore, commonly used GIS libraries, such as GDAL/OGR and GeoPandas, fully support reading semi-structured GIS formats, so it is often used in spatial analysis workflows of all kinds, not just mapping. Table 1 lists a few well-known types of semi-structured GIS data along with their common applications.

Table 1. Examples of Semi-Structured GIS Data Sets with Common Applications.

Type	Format	Contents	Applications
OpenStreetMap (OSM)	XML, GeoJSON	Geographic features such as roads, buildings, points of interest, plus user-defined tags	navigation, urban planning, disaster response
USGS Earthquake Catalog	GeoJSON	Earthquake event data with attributes such as magnitude, depth, and location	seismic monitoring and risk assessment
NOAA Global Forecast System (GFS) Weather Data	GRIB (binary format), JSON	Weather forecast data with attributes such as temperature, wind speed, and precipitation probabilities	meteorological forecasting, climate research, and disaster preparedness
Natural Earth	GeoJSON	Cultural and physcal datasets for cartography	mapmaking and geographic visualization

It is important to note that semi-structured formats, such as GeoJSON, can encode both object and network models. For example, OpenStreetMap data represents transportation features as a connected network graph. The TopoJSON extension of GeoJSON further encodes topological relations between objects.

3. Unstructured Geospatial Data

Unlike structured and semi-structured data, unstructured data is an ad hoc category that includes any data that has no standard formatting. Most data (by some estimates 80-90%) on the web is unstructured and includes social media posts, emails, newspaper articles, documents, web pages, multimedia images, videos, sound files, sensor observations, and more (Harbart, 2021).

Unstructured data can be used for geospatial analysis when there is geographic metadata or the data itself contains geospatial information. Streams from in-situ sensor networks (GPS trackers, weather stations, and other sensors) produce unstructured spatial logs. These data can be structured into time-stamped point series or gridded fields for analysis. Remote sensing data, although explicitly spatial, can be unstructured, e.g., LiDAR (Light Detection and Ranging) unordered 3D point clouds, and require specialized tools to add structure. Other unstructured geographic data have only minimal spatial content, e.g. point metadata, or no explicit spatial content at all.

Unstructured data forces an analyst to distinguish between spatial and geographic representation, illustrating that computational geographical and spatial analysis, while connected in many ways, are not interchangeable and often require distinct approaches.

3.1 Unstructured Geospatial Imagery

While satellite and aerial images are often treated as structured raster data, many forms of photographic imagery are unstructured in nature, especially when captured through ad hoc or crowdsourced methods. Imagery can be multispectral and outputs from geospatial workflows include DEMs and 3D reconstructions.

In theory, given the location of a camera and its orientation, any image taken on Earth can be mapped to a three-dimensional spatial coordinate system. In practice, metadata typically includes only a single point (latitude and longitude) with no orientation, limiting its spatial precision. The value of such imagery increases when used in aggregate, for

example, crowdsourced photos taken by diferent people in the same location, or sequences collected systematically by drones. These can be processed using Structure from Motion (SfM) (Schonberger & Frahm, 2016), a technique that reconstructs 3D geometry by matching features across multiple overlapping images. SfM allows the generation of dense point clouds, terrain surfaces, and orthophotos without requiring specialized sensors. Table 2 summarizes some of the geospatial data products that can be derived from imagery using SfM techniques.

Table 2. Examples of Geospatial Data Derived from Images using Structure from Motion (SfM).

Geospatial Data Generated	Description	Format Generated
Digital Elevation Models (DEMs)	3D terrain models from aerial drone imagery for topographic analysis	GeoTIFF, point cloud
3D Models	models of buildings and environments for urban planning, cultural heritage, and archaeology	OBJ, CityGML
Orthomosaics	high-resolution stitched photographs	GeoTIFF, BigTIFF, JPEG2000, EWC
Vegetation and Land Cover	classification for measuring vegetation and analyzing land-use patterns	GeoTIFF, other raster formats

A fast-changing set of machine learning techniques are now used to analyze raw imagery in geospatial workflows. Convolutional Neural Networks (CNNs) are increasingly used to classify regions, detect objects, and segment features from highresolution aerial or satellite data (Maggiori et al. 2016). These tools and related techniques are discussed in more detail in Section 4.

3.2 Unstructured Geospatial Text

While image data allows us to observe the changing state of the natural and urban environment, unstructured textual data is where we can best access geographic knowledge filtered through an individual experiential, social, or cultural lens (Adams & McKenzie 2013). Consequently, unstructured textual data is uniquely suited for answering certain kinds of social and environmental research questions (Purves et al. 2022).

Just as image analysis has improved immensely in the last decade with the development of new computer vision models, analysis of textual data has advanced alongside innovations in natural language processing (NLP). An NLP technique, named entity recognition (NER), is used to identify place names and references to geospatial entities and events in text (Lample et al. 2016). Transformer-based language models (e.g. BERT, GPT) have achieved remarkable performance on named-entity recognition and related tasks.

A common way to spatialize text is to use a gazetteer (a spatial database of well-known places) to link entities in the text to location information identified via NER. For instance, a transformer-based geoparsing pipeline outperforms traditional methods in identifying and geocoding toponyms (Berragan et al. 2022). Many other NLP methods can be applied to geospatial text. Table 3 shows some geographic applications where unstructured text is useful alongside the NLP techniques commonly used.

Table 3. Geospatial Application Areas that Rely on Analysis of Unstructured Data. Some indicative data sources and computational techniques are shown.

Application Area	Textual Data Sources	Techniques
Disaster response and humanitarian aid	social media, news articles, reports from government and non-profit organizations	NER, geocoding, sentiment analysis, temporal tracking
Journalism	social media, news articles, blogs, leaked documents	NER, topic modeling, cross-source validation, network analysis
Market analysis	customer reviews, social media, industry reports	NER, sentiment analysis
Public health	social media, health reports, electronic health records	NER, geocoding, predictive modeling, trend analysis
Environmental monitoring	scientific papers, citizen science data, environmental reports	NER, geocoding, trend analysis
History and cultural analytics	historical manuscripts, archaeological reports, historical maps	Historical NER, knowledge graphs, geospatial linking
Urban planning	social media, planning reports, real estate listings	NER, geocoding, sentiment analysis, predictive modeling

Figure 1 shows a sample workflow explaining how unstructured social media data could be processed and used in a disaster response scenario. The first step involves collecting relevant data from real-time social media feeds and news sources. This can be done by filtering on keywords or hashtags or filtering based on optional regional information from the data provider. Because unstructured text data includes many features that are noisy, a cleaning step is applied to remove unnecessary information. After this, NER is used to identify named places in the posts which are then geocoded to coordinates making it usable for spatial analysis. Using a machine learning classifier, the posts that match the relevant sentiment or topics can be identified. Hotspots are identified by applying spatial analysis, e.g. clustering, and then visualized for decision support.

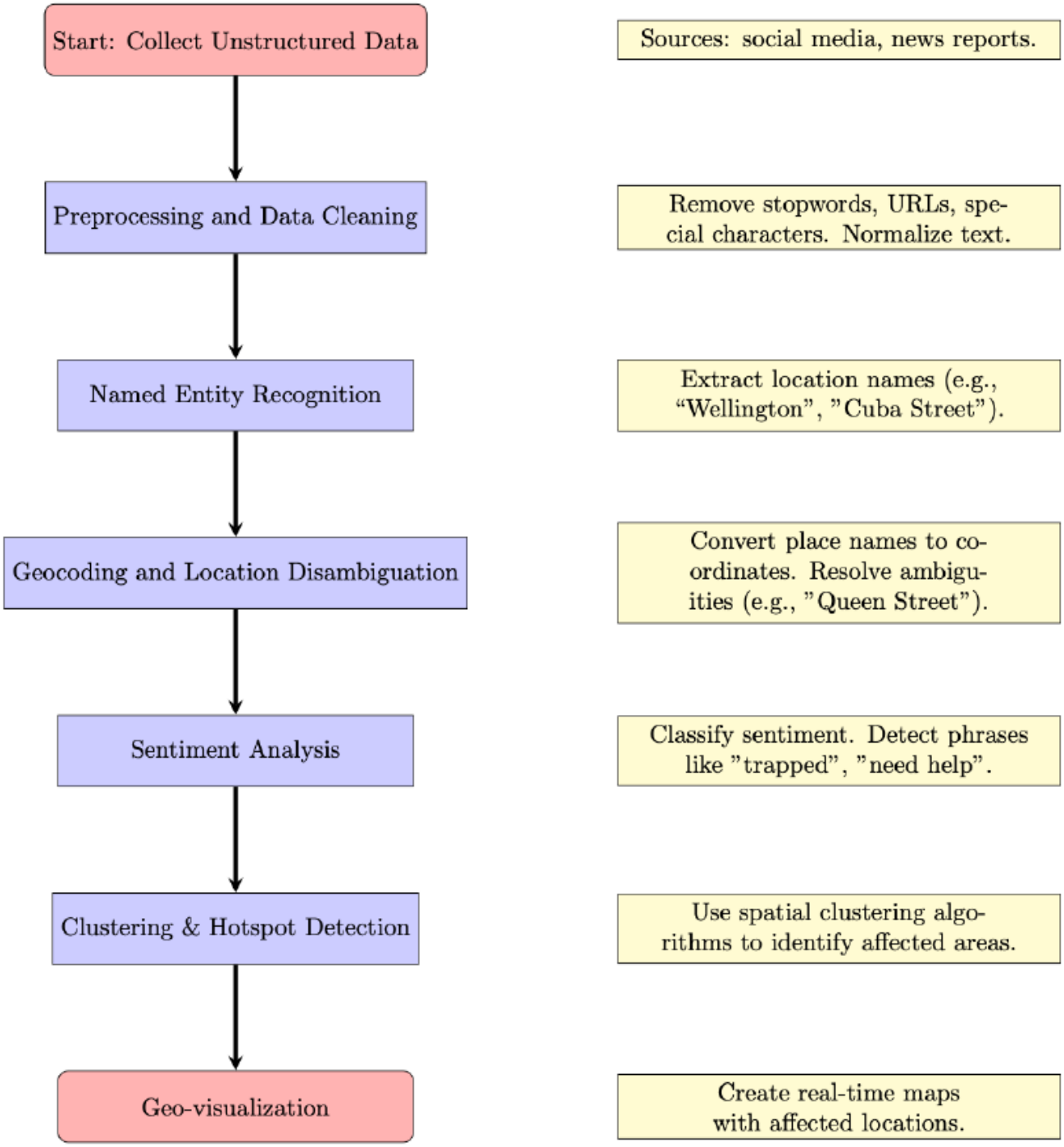

Figure 1. A Sample Workflow to Analyze Unstructured Data for Disaster Response. Source: author.

3.3 Multimodal Data

Humans experience the world and develop conceptualizations of space and place through the integration of various modalities made up of visual, aural, language, and sensorimotor information. The data we share about the world mirrors that diversity of experience; e.g., we might share a description of place attached to an image on social media. The methods and data described above treat these diferent modalities as distinct types of geographic

data, however in reality some of the richest digital geographic information available is multi-modal. The techniques and methods used to analyze these data are as varied as the forms these data take. In the future, we can expect that multimodal AI tools which flexibly adapt to many diferent kinds of data will play an increasingly important role in how we analyze unstructured geographic data (Ngiam et al. 2011; Kiros et al. 2014).

4. Machine Learning and AI for Spatial Data

Although machine learning methods have long been applied to spatial data, recent innovations in AI have revolutionized semi-/unstructured spatial data analysis. AI methods are now essential for extracting structure from the growing volume of semiand unstructured spatial data. Across imagery, text, point clouds, and sensor logs, machine learning workflows allow researchers to automate classification, prediction, and pattern recognition with high spatial and temporal resolution.

4.1 Image-based Spatial Learning

Deep learning methods, particularly CNN architectures such as U-Net and ResNet, are widely used in geospatial image analysis (Ronneberger et al. 2015; He et al. 2016). These models classify land cover and segment features in raster images. Libraries like Raster Vision (https://rastervision.io/) and TorchGeo (https://www.osgeo.org/projects/torchgeo/) support workflows for semantic segmentation and object detection on georeferenced imagery. Massive, commercial pre-trained image foundation models of Earth imagery are also available for researchers to leverage (Alpha Earth Foundations Team, 2025).

4.2 3D Point Clouds and Surface Models

The Point Data Abstraction Library (PDAL) is an open source C++/Python library for translating, filtering / denoising, and classifying point clouds (https://pdal.io/). Neural networks such as PointNet (Qi et al. 2017) can ingest raw point clouds for classification and segmentation. These methods have been used extensively for forestry analysis, terrain mapping, modeling built environments, and autonomous driving.

4.3 Textual Geospatial Data

As discussed in Section 3.2, natural language text is a major source of unstructured spatial information. Recent advances in NLP, especially commercial large language transformer-based models such as OpenAI ChatGPT, Google Gemini, and Anthropic Claude are increasingly being integrated into geospatial workflows. Pretrained language models can now be fine-tuned for geospatial applications such as disaster response or cultural analytics; however, ongoing concerns about geographic bias in large models persist (Dunn et al. 2024).

4.4 Sensor Networks and Spatiotemporal Learning

Sensor networks generate high-velocity streams of semi-structured spatial data. To analyze trends and make predictions, sequence models such as temporal convolutional networks (TCNs) are used (Lea et al. 2016). These models capture both spatial autocorrelation and temporal dependencies. Graph neural networks (GNNs) are also being applied to spatial

networks (e.g. road connectivity, social flows), ofering new ways to model relationships and flows in geographic space (Wu et al. 2020).

4.5 Multimodal and Hybrid Models

Multimodal models that combine image and language analysis with geospatial reasoning is a growing area. For example, CLIP (Radford et al., 2021) links image features with textual descriptions, enabling zero-shot classification of satellite imagery. Recent eforts to apply large foundation models (e.g. GPT-4, Gemini) to geospatial questions point to how we might build systems that can integrate spatial data, text, and imagery seamlessly, however this remains an open area of research (Ji et al. 2025).

5. Conclusion

In this chapter we presented some examples of these kinds of data and the research applications for which semi- and unstructured data are used. Semi- and unstructured spatial data ofer rich insights but require multidisciplinary modeling techniques. The intersection of GIS, remote sensing, statistics, and AI is rapidly advancing. Bespoke analysis methods and the application of advanced computational techniques are often required; it is likely that continuing advancements in AI tooling will help make these data sets more accessible and usable for analysts working across a variety of research domains.

References

Adams, B., & McKenzie, G. (2012). Inferring thematic places from spatially referenced natural language descriptions. In Crowdsourcing geographic knowledge: Volunteered geographic information (VGI) in theory and practice (pp. 201-221). Springer.

Alpha Earth Foundations Team, The. (2025). AlphaEarth Foundations helps map our planet in unprecedented detail. Accessed August 20, 2025.

Berragan, C., Singleton, A., Calafiore, A., & Morley, J. (2023). Transformer based named entity recognition for place name extraction from unstructured text. International Journal of Geographical Information Science. 37(4), 747–766.

Charles, R. Q., Su, H., Kaichun, M. and Guibas, L. J. (2017). "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 77-85.

Dunn, J., Adams, B., & Madabushi, H. T. (2024). Pre-Trained Language Models Represent Some Geographic Populations Better than Others. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 12966-12976).

Harbart, T. (2021, February 1). Tapping the power of unstructured data. Blog posting, MIT School of Management.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In

2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778.

Ji, Y., Gao, S., Nie, Y., Majić, I., & Janowicz, K. (2025). Foundation models for geospatial reasoning: assessing the capabilities of large language models in understanding geometries and topological spatial relations. International Journal of Geographical Information Science, 1-38.

Kiros, R., Salakhutdinov, R. & Zemel, R. (2014). Multimodal Neural Language Models. Proceedings of the 31st International Conference on Machine Learning, in Proceedings of Machine Learning Research 32(2):595-603.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 260–270, San Diego, California. Association for Computational Linguistics.

Lea, C., Vidal, R., Reiter, A., Hager, G.D. (2016). Temporal Convolutional Networks: A Unified Approach to Action Segmentation. In: Hua, G., Jégou, H. (eds) Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science(), vol 9915. Springer, Cham.

Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017). Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification. IEEE Transactions on geoscience and remote sensing, 55(2), 645-657.

Mooney, P., & Minghini, M. (2017). A review of OpenStreetMap data. In G. Foody, L. See, S. Fritz, P. Mooney, A.-M. Olteanu-Raimond, C. C. Fonte, & V. Antoniou (Eds.), Mapping and the Citizen Sensor (pp. 37–59). Ubiquity Press.

Mueller, C. S. (2019). Earthquake catalogs for the USGS national seismic hazard maps. Seismological Research Letters, 90(1), 251-261.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal Deep Learning. In Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, (Vol. 11, pp. 689-696).

Purves, R., Koblet, O., & Adams, B. (Eds.) (2022). Unlocking Environmental Narratives: Towards Understanding Human Environment Interactions through Computational Text Analysis. London: Ubiquity Press.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ..., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748-8763). PmLR.

Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical

Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham.

Schönberger, J. L. and J. -M. Frahm, J.-M. (2016). "Structure-from-Motion Revisited," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 4104-4113.

Wu, N., Zhao, X. W., Wang, J., and Pan, D. (2020). Learning Effective Road Network Representation with Hierarchical Graph Neural Networks. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20). Association for Computing Machinery, New York, NY, USA, 6–14.