AGILE: GIScience Series, 7, 11, 2026. https://doi.org/10.5194/agile-giss-7-11-2026
Proceedings of the 29th AGILE Conference on Geographic Information Science, 16–19 June 2026.
Eds.: Evelyn Uuemaa, Alexander Kmoch, Holger Virro, and Simon Scheider
This contribution underwent peer review based on a full paper submission.

Exploring urban polygonal representation learning for complex footprint groups

Luisa Lo Presti [1] and Peter Mooney [2]

[1]Hamilton Institute, Maynooth University, Ireland
[2]Department of Computer Science, Maynooth University, Ireland

Correspondence: Luisa Lo Presti (luisa.lopresti.2024@mumail.ie)

Abstract. Urban representation learning has traditionally focused on human mobility patterns and space functionality derived from points of interest. While such approaches provide a powerful way of understanding interactions between people and the urban environment, they overlook the impact of the physical urban configuration. When urban form is considered, its representation is commonly achieved using rasterization or abstractions, such as zonal aggregation and graph constructions, leading to information loss, sensitivity to design choices, and limited generalizability. Borrowing from signal processing, spectral approaches have been developed to encode polygonal geometries directly. Yet, their implications for urban systems remain largely under-explored. In this work, we address this gap by investigating the application of this methodology in complex urban scenarios, focusing on the unsupervised learning of compact embeddings for groups of building footprints. Using a reproducible workflow without rasterization or graph abstraction, the resulting embeddings are designed to capture the structural configuration of building clusters for urban form analysis beyond handcrafted features. The learned embeddings are analyzed using qualitative and quantitative evaluation methods. Our results demonstrate the novel potential of spectral approaches to learn generalizable, geometric urban embeddings and highlight their applicability for a wide range of urban analysis tasks.

Submission Type. Analysis.

BoK Concepts. [GC1] Geocomputation and complex systems; [GC3] Artificial intelligence (AI) in EO and GI.

Keywords. Urban Representation, Building Footprints, Shape and Topology, Non-Uniform Fourier Transform, Latent Space Analysis.

1 Introduction and Context

The modern world is characterized by a significant degree of urbanization, often underestimated by national statistics due to unharmonised definitions of *"urban areas"* (United Nations, Department of Economic and Social Affairs, 2025). As cities continue to steadily grow, the United Nations notes that built-up areas are expanding faster than population growth worldwide, and highlights how *GeoAI* can be a transformative tool to enhance accuracy, efficiency, and capabilities (United Nations, 2026). These developments have led to a proliferation of studies on the urban environment, with many focusing on urban functions, mobility, and Points of Interest (POIs) (Zhang et al., 2024; De Sabbata et al., 2023; Niu and Silva, 2021; Huang et al., 2018; Sun et al., 2024; Qin et al., 2024). These studies learn compact, machine-readable embeddings of different urban entities (e.g., neighbourhoods, roads infrastructures, POIs) so that complex urban phenomena can be analyzed, predicted, or compared. To this end, human trajectory data and POIs often play a complementary role: the former enable dynamic insights into mobility patterns, connectivity, accessibility, human activities and land use, while the latter provide static information about the function and semantic of the built environment. Thus, POIs represent the design of the urban space, while mobility data the interaction of people with such design. Together they enable more nuanced understanding of urban dynamics to inform urban planning strategies. For instance, Niu and Silva (2021) use vectors of POI classes to calculate similarity scores, forming clusters of areas characterized by shared uses, uncovering latent functions. While this approach effectively captures the functional composition of the built environment, it does not explicitly account for the geometric configuration of the urban form, focusing instead on attributes derived from POI distributions. The role of spatial configuration in shaping how these functions are experienced is left at the margin of most studies. However, street arrangements, city

blocks, and building configurations influence accessibility, connectivity, and human movement, which cannot be fully inferred from functional data alone. Consequently, areas with similar functional profiles may exhibit substantially different urban dynamics depending on their morphological characteristics. Moreover, urban form shapes environmental outcomes, such as microclimate, energy use and land use intensity; ignoring geometric layouts implies overlooking potentially critical variations in how urban systems operate. A different approach is employed by Zhang et al. (2024), who leverage unsupervised graph representation learning by integrating spatial, temporal and transitional views of human trajectories. This approach allows for a rich understanding of urban dynamics and produces embeddings that generalize well across different urban tasks. However, such study relies heavily on human activity data, which limits scalability due to the high cost and restricted coverage of the data collected.

In this paper, we introduce a modular component that can be integrated into multimodal pipelines to explicitly capture the effects of the 2D building configurations, augmenting existing representations of urban environments with morphological information and improving their ability to reflect geometric structure. This represents a first step towards enabling the characterization of urban areas not only by functional profiles, but also by physical layouts, allowing models to distinguish between areas with similar functions but different morphologies, and supporting downstream tasks related to urban form.

Geometric Representation Advances

Only a few studies have focused on modelling the physical representation of the urban built environment (Li et al., 2023). Yet, when urban form and structure are incorporated, this is frequently achieved using zonal statistics and spatial aggregation, leading to a loss of spatial granularity and of local heterogeneity, and high sensitivity to the selected spatial partition (Wang et al., 2025). In addition, due to the wide availability of raster-based methods, geometries are frequently rasterized (Balsebre et al., 2024; Choudhury et al., 2026), which inevitably introduces boundary approximation errors, loss of precision and of topological properties, and a fixed resolution that prevents multi-scale analyses (Van't Veer et al., 2019). In response to these limitations, previous work developing urban representation from the built environment have often relied on graph-based structures (Yan et al., 2021; Liu et al., 2025; He et al., 2018; Bei et al., 2019) or convolutional neural networks (Li et al., 2023; Van't Veer et al., 2019). Graph neural networks (GNNs) are particularly well-suited for tasks that require explicit spatial and functional relationships, and they have proven capable of accommodating irregular urban layouts and heterogeneous features. Nevertheless, their performances are highly sensitive to non-trivial construction choices (e.g., node definition, edge criteria, scales, etc.), often requiring city-specific knowledge about urban morphology that hinder generalizability across different geographies. On the other hand, methods based on convolutional neural networks (CNNs) focus on local features, which results in limited awareness of the global structure and often disrupts the geometrical topology. Moreover, both graph-based and CNN-based representations of urban shapes are typically suitable for simple polygons, but can perform poorly on complex building geometries, such as polygons with interiors or multi-part geometries. This has the effect of leading to a loss of morphological detail (Yan et al., 2021; Van't Veer et al., 2019). Such limitation derives from the fact that shapes are not treated as continuous geometric regions. However, recent work in geometric representation learning has developed new end-to-end approaches, allowing researchers to model vector data directly, with several advantages for urban forms and topology studies (Yan et al., 2019; Mai et al., 2022).

Building on the work done on a deep differentiable simplex layer for learning geometric signals (Jiang et al., 2019), Mai et al. (2022) introduced an approach based on the Non-Uniform Fourier Transform (NUFT) to directly model polygonal geometries, demonstrating advantages over traditional convolutional methods. While their evaluation focused on a vectorized version of the MNIST dataset (Lecun et al., 1998), the overall methodology in Mai et al. (2022) is particularly appealing for urban systems, since it allows vector-based representations that preserve shape and topology. This has the advantage of preventing both losses and aliasing derived from rasterization and abstractions, and of avoiding the extensive preprocessing associated with graph-based approaches. NUFT-based methods have the potential to better represent complex urban forms (including simple polygons, polygons with interiors, and multi-part geometries), preventing overfitting to local details, and enhancing generalizability and global shape representation.

2 Our Approach and Contributions

In this work, we build upon the current literature, adapting the mentioned NUFT-based approach to real-life scenarios in urban representation learning. Specifically, we focus on learning a compact representation of building groups, defined as clusters of building footprints within the same urban boundaries (i.e., roads and waterways). The learned embeddings are designed to capture the different geometric configurations of the studied groups, going beyond traditional handcrafted features and thus moving toward a general-purpose representation for urban analysis. Such representations support a variety of layout-dependent tasks, including similarity analysis and retrieval, topological classification, and comparative morphological studies, but also contribute to land use and

land cover classification, urban function approximations and regeneration analysis. The specific contributions of this research work are outlined as follows:

1. We develop a reproducible workflow to build coherent city blocks and building groups extracted from OpenStreetMap (OSM) building footprints. This extends the NUFT-based geometric representation approach to complex urban regions in an unsupervised setting. At the time of writing, we believe that our work is one of the very first research works to use complex building geometries for urban representation learning.

2. We provide an in-depth evaluation of our methodology, analysing latent space structures and downstream task performance to assess strengths and limitations of our approach in urban contexts.

3. We outline concrete pathways to improve urban representation learning through combined frequency-based and deep learning techniques, aiming towards more generalizable and geometry-aware urban embeddings.

These contributions demonstrate the novelty and potential of using complex building geometries for urban representation learning. Our approach provides a reproducible, unsupervised framework that captures rich structural patterns in urban areas, offering insights into the latent organization of cities. This work represents an initial step toward developing more generalizable, geometry-aware urban embeddings.

The remainder of this paper is organized as follows. In Section 3, we discuss the underlying methodology, including terminology, dataset, proposed model and training strategy, followed by the introduction of a case study. Results and applications are discussed in Section 4. Finally, in Section 5, we draw conclusion and outline strategies for improved geometric representation learning.

3 Methodology and Setup

This section presents the methodology underlying our study. In Section 3.1, we define some of the relevant terms; in Section 3.2, we describe the dataset; in Section 3.3, we introduce the proposed model; in Section 3.4, we outline our training choices; finally, in Section 3.5, we illustrate the selected case study.

3.1 Methodology: Definitions

To ease reader comprehension of the terminology used, we propose a few simple definitions and state our exploratory problem.

Definition 1: building footprint. A building footprint is defined as a 2D polygonal geometry representing a building outline. A building footprint can consist of one or multiple separate components (i.e., multi-polygons). Each component is defined by an exterior boundary and optionally one or more interior boundaries, representing holes in the geometry. All boundaries are defined as a sequence of coordinates forming a closed ring.

Definition 2: building block. A building block is defined as a set of building footprints obtained by partitioning the geographical space according to the natural boundaries imposed by roads and waterways.

Definition 3: building group or building cluster. A building group or building cluster is a subset of a building block. This is derived by its partitioning through buffer-analysis. Two building footprints within the same building block are considered to belong to the same building group if the buffers associated with their footprints intersect. Each building block contains at least one building group.

Problem statement. Given a set of building groups $G = \{g_1, g_2, ...\}$, the objective is to learn an embedding function $f : G \rightarrow R^d$ in an unsupervised manner, such that each group g_i is mapped to a compact vector representation $e_i = f(g_i)$. All embeddings $e_i \in E = \{e_1, e_2, ...\}$ share a common dimensionality d and are learned in an unsupervised and contrastive setting. They are intended to preserve the salient geometric and structural properties of the corresponding building group.

3.2 Methodology: Dataset Description

Our approach to urban representation learning utilizes groups of building footprints obtained from OSM as input data. Based on cartographic studies and building generalization theory (Basaraner and Selcuk, 2008), we first apply the openly accessible generalization algorithm described in Lo Presti and Mooney (2025) to roads and waterways in order to derive natural boundaries. These generalized elements represent the first partition of the urban space into building blocks. These blocks are partitioned further into building groups via buffer-based cluster analysis. Similar proximity-based approaches have been proved to be the most effective for creating accurate grouping (Deng et al., 2018) and have been widely used in the literature (Yan et al., 2019; He et al., 2018). This approach generates a variety of different geometries, which, depending on shape complexity and footprint density, can be simple polygons, polygons with internal boundaries, or multi-polygons.
In order to prepare our building groups data for the modelling stage, we perform several data preparation steps, including geometry normalization. This involves centring the geometry bounding box to the origin, scaling to unit space, and translating the result into the positive coordinates. Normalized building groups are used to construct 2-simplex meshes by adding an auxiliary node in the origin. The 2-simplex mesh is defined as $S^2 =$

$\{S_n^2\}_{n=1}^N = (V, E, D)$, where N is the total number of building groups, V is the matrix of vertices, E is the matrix of edges (indices of connected vertices), and D is the density matrix (set to a constant 1). Representing complex building groups via V and E matrices enables direct learning. It is important to note that internal boundaries should be encoded in a clockwise orientation, while external boundaries should be encoded counter-clockwise. This ensures topology preservation by attributing correctly positive and negative signs to different area portions. For further details on this, the reader is referred to Mai et al. (2022).

3.3 Methodology: Model

The proposed model takes as input the 2-simplex meshes obtained in Section 3.2, and performs a NUFT over them using a geometric grid. The result of the spectral transformation is passed to a Multi-Layer Perceptron (MLP), which represents the learnable part of the model. We adopt a MLP here as a baseline architecture to evaluate the effectiveness of the methodology, leaving exploration of more complex architectures for future work. The MLP is trained to produce meaningful embeddings that discriminate between different geometries. To this end, the spectral features should ease the learning of salient characteristics, enabling different downstream tasks. The MLP consists of five-layers.

1. The first layer takes as input the NUFT spectral features (2176 dimensions) and applies a linear projection to 1024 dimensions, followed by ReLU activation, dropout with rate 0.1, and layer normalization.

2. The second and third layers are represented by two residual layers (Mac Aodha et al., 2019; He et al., 2016), each preserving the 1024-dimensionality of the latent space and employing the same ReLU activation, dropout rate and layer normalization.

3. A subsequent compression layer reduces the feature dimensionality from 1024 to 256, again using ReLU activation, layer normalization, and 0.1 dropout rate.

4. Finally, an output layer maps the 256-dimensional representation to 128 dimensions using a linear transformation with identity activation.

The overall architecture is illustrated in Fig. 1. The approach has several interesting and novel properties relevant to geometric representation of urban systems. The encoder representation is robust to affinity transformations, exhibiting the following properties: (1) loop-origin invariance, ensuring that the representation of a footprint remains consistent independently of the first vertex encoded; (2) trivial vertices invariance, preventing the addition of vertices that do not alter the footprint shape from influencing the learned embeddings;

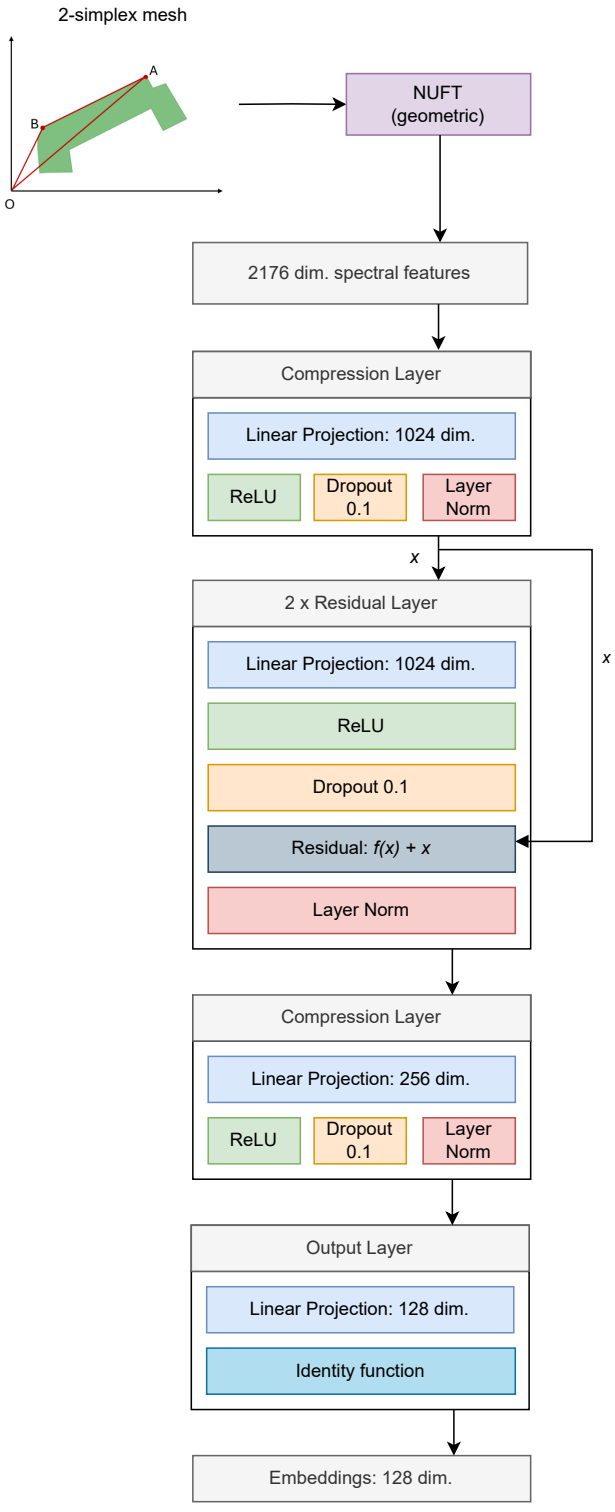

Figure 1. Diagram of the proposed model architecture.

and (3) part permutation invariance, allowing complex multi-part structures, like apartment blocks or university departments, to be represented consistently regardless of the order in which individual buildings are processed. This enables the explicit learning of topological features. Moreover, the simplex-based decomposition allows an end-to-end approach, avoiding extensive feature

engineering and computation of zonal statistics, such as building densities per spatial unit, which hinders generalizability and can cause information loss.

3.4 Methodology: Training

Differently from most ready-to-use datasets, real-life data often come unlabelled, thus producing accurate labels to train a supervised model is a costly and long procedure. Therefore, we rely on an unsupervised setting to extract information from urban geometries. In our experiment, we learn over building groups' simplices using an InfoNCE (Information Noise-Contrastive Estimation) contrastive loss (Van den Oord et al., 2019). For each complex geometry (i.e., each building group) in a batch, we generate the corresponding augmented version, serving as positive pair, while the other geometries in the batch represent the negative pairs. The augmentation used to obtain positive pairs rely on affine transformations that do not significantly alter the topology. Such augmentations include rotation, translation, flipping, and light random Gaussian noise on vertices coordinates. In this way, we train the model to recognise geometric patterns while discarding subtle differences. The model is set to train for 300 epochs, but an early stopping with 15 epochs of patience is enforced. The selected optimizer is Adam, which combines momentum and adaptive learning rates to accelerate convergence and stabilise training. It is used with a learning rate of 10^{-4} and an L2 penalty, thereby controlling model complexity and improving generalization performance.
All parameters are reported in the associated GitHub repository at the link provided in the ***Data and Software Availability*** declaration at the end of this paper.

3.5 Methodology: Case Study

We select Milan (Italy) as a case study. As many European cities, Milan develops around its centre (Boeing, 2021), with high morphological and structural variability (Taubenböck et al., 2020). The diversity of building footprints and agglomerations make Milan a solid test-bed for our approach, enabling the evaluation of the model on diversified geometries. Nevertheless, to demonstrate generalizability, future work will apply the approach to diverse urban configurations, including coarse and fine-grained grids, unplanned layouts, and organic forms.
We retrieve all building geometries from OSM using the *osmnx* (Boeing, 2017) Python library, and process them as described in Section 3.2. To avoid data leakage due to spatial autocorrelation, we divide our data into *train*, *validation* and *test* sets using contiguous spatial partitions (Roberts et al., 2017). In particular, we build the convex hull encompassing our data and divide it in 20 sections of equal areas, each representing ~5% of the total convex hull. We assign contiguous sections encompassing 70% of the convex hull to the *train* set, select sections around the training data for validation, and keep the most detached sections for testing. Both *validation* and *test* are assigned 15% of the total hull area. Through this methodology, we aim to ensure a balanced representation of both central and peripheral geometries across all datasets, while also guaranteeing that predictions on the *test* set are reliable and not driven by the physical proximity of samples observed in the training set. This partition results in 8816 training building groups (~65.5% of the city data), 1842 validation groups (~13.7%), and 2805 test groups (~20.8%). The corresponding split for the case study city of Milan is shown in Fig. 2.

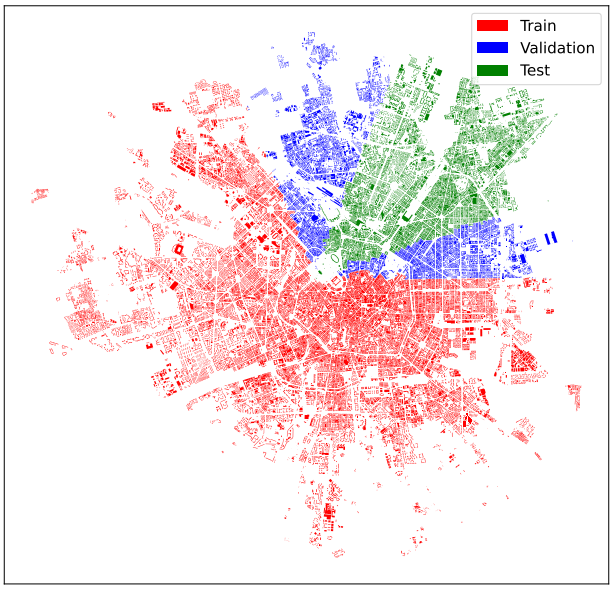

Figure 2. Building clusters obtained for the city of Milan (Italy), partitioned into *train*, *validation* and *test* sets.

4 Evaluation of Geometric Urban Embeddings

Obtaining general-purpose geometric representation of the built environment is of great interest, as it enables to study different urban phenomena using a single geometric representation. The spatial arrangement of buildings plays a fundamental role in shaping human activities, environmental processes, population density, and land use. Thus, urban planning, development, and regeneration strategies are strongly influenced by the geometric layout of the built environment. In addition, socio-economic activities and accessibility are closely linked to building configurations. Beyond these applications, such representations are also valuable for morphological analysis and geometric retrieval tasks. Moreover, general-purpose geometric embeddings have the potential to overcome the limitations of conventional descriptors, such as:

1. incomplete shape information,

2. the requirement for extensive domain expertise and manual tuning,

3. lack of generalizability across geographic areas,

4. lack of generalizability across different tasks in the same geographical area.

To explore the strengths and limitations of the proposed model and analyze potential pathways forward, we evaluate the goodness of the test embeddings using separate approaches, considering both qualitative and quantitative perspectives. In Section 4.1, we apply dimensionality reduction on the test embeddings and visually explore clusters and their geometric characteristics. Following this, in Section 4.2, we propose a quantitative evaluation of the embeddings via land cover binary classification task. Robust qualitative and quantitative results would imply that the learned embeddings capture meaningful and transferable geometric characteristics of urban form. Such results would support the usage of these representations for the above mentioned downstream tasks and their extensions to other urban contexts.

4.1 Qualitative Evaluation: Dimensionality Reduction

To qualitatively explore the latent space, we apply the t-distributed Stochastic Neighbors Embedding (t-SNE) algorithm (Van der Maaten and Hinton, 2008) on the embeddings. Given the high dimensionality, we firstly extract the principal components preserving 95% of the variance, and then apply t-SNE to these components. For visual purposes, we set t-SNE components to two and evaluate the presence of clusters in the obtained latent space. The high-dimensional latent features are projected into two dimensions, so they can be plotted on a 2D scatter plot. Observing the points forming distinct clusters provides an indication that the model has learned meaningful separations between samples. We apply a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996) on the 2D latent space and consider the resulting clusters. The results are shown in Fig. 3. DBSCAN performs clustering based on the distance between observations in the latent space; moreover, it has the ability to identify "noise", hence observations that are sparse and scattered in space and are not assigned to any specific cluster. DBSCAN noise likely corresponds either to geometries that our model fails to represent properly or to very peculiar and rare shapes that differ significantly from others in the dataset. In Fig. 3, noisy observations are portrayed in blue color and are scattered across the observation space.

Observing the t-SNE result, we can easily recognize a few clusters in the periphery of the plot that are significantly separated from the core. This separation is desirable as it shows embeddings with high discriminative power. We observe that well-separated clusters are characterized by relatively simple shapes, i.e., mostly single-part geometries, with or without holes. This implies the model has a strong capability to handle such geometries in urban configurations. Conversely, the core of the plot shows a higher density of observations with reduced separation between clusters, indicating that for more complex polygonal structures the model has learned a gradient of representations rather than clearly defined boundaries. Examining these dense regions reveals that neighbouring geometries share meaningful structural characteristics, suggesting that the embeddings captures geometric similarity and relevant morphological features. The reduced separation among clusters is associated with the increased complexity of building groups, which exhibit distinctive or unusual patterns. As an example of complex cluster, we report cluster 16 at the left-hand side of Fig. 3. The displayed geometries show similar patterns in their outline shape and elongation. However, their increased complexity and irregularities makes it challenging to form well-separated embeddings.

Nevertheless, inspection of the clusters shows that groups are characterized by similar shapes, elongations, and complexities, remarking the capability to capture geometric patterns. For instance, cluster 85 displays blocks composed of approximately rectangular shapes arranged in similar relative positions; cluster 114 contains L-shaped polygons; cluster 45 is composed of square-like shapes. Additional examples of clusters are available in the GitHub repository.

The visual inspection demonstrates that geometries within the same clusters share key structural features, confirming the capabilities of the proposed encoder. Results are satisfactory given the simplicity of the model used (i.e., an MLP), which performs reasonably well. Architectures more suitable for complex geometrical patterns will be explored in future work, and potential directions are highlighted in Section 5.

4.2 Quantitative Evaluation: Land Cover Classification

To quantitatively evaluate the model, we use the learned representations to perform land cover classification. Land cover and building use classification are well-established tasks in GIS (Li et al., 2023; Balsebre et al., 2024; Ren et al., 2024; Hecht et al., 2015; Oostwegel et al., 2025; Atwal et al., 2022), and therefore provide a useful benchmark for assessing the quality of deep embeddings.

To this end, we use the trained encoder to extract embeddings for building blocks (definition 2), and feed these representations into simple classifiers. Although the encoder is trained on building clusters (definition 3), it generalizes effectively to more complex polygonal geometries (i.e., building blocks), which are more appropriate for land cover classification. Training the encoder on clusters allows us to expand the size of the training set; however, blocks represent a more natural

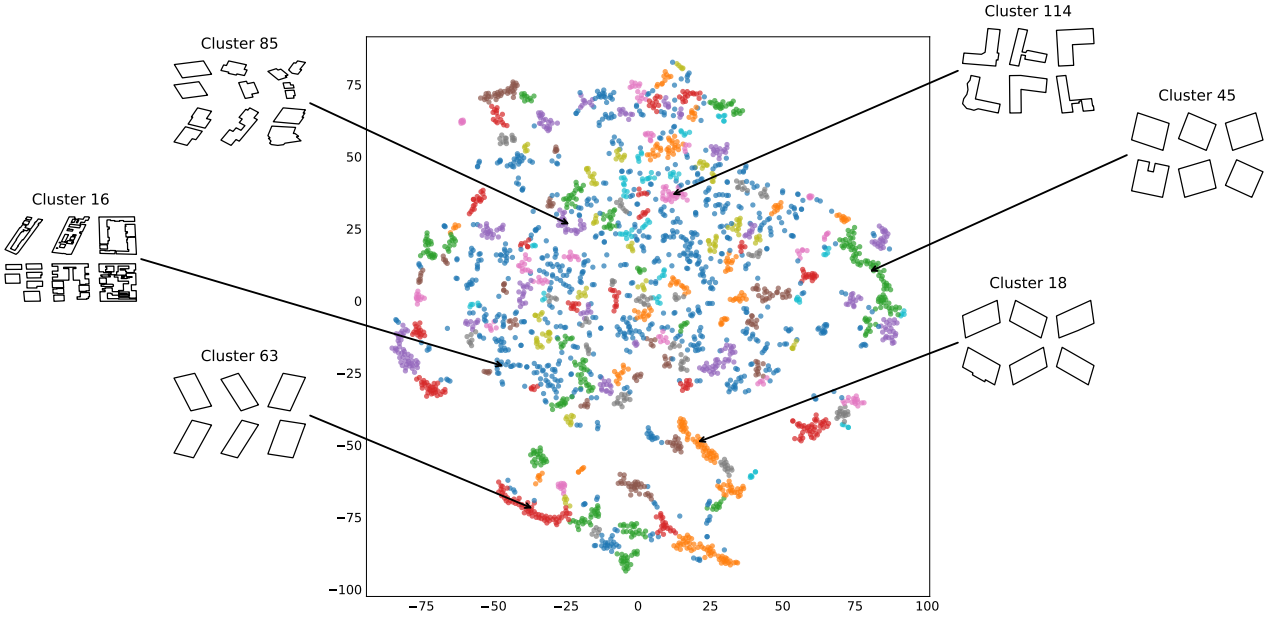

Figure 3. Results of the t-SNE algorithm performed on the *test* embeddings. Observations are colored according to their corresponding DBSCAN cluster. For illustration purposes, six clusters are shown, each with six representative geometries.

unit for land cover classification, as the task depends not only on form, but also on spatial relations. While our encoder focuses primarily on geometric form, the partition of the city into blocks inherently embeds spatial context into the data. By treating all buildings within the same block as a single observation, the model captures implicit relational information through their collective structure, as individual footprints provide mutual spatial context for one another.

CORINE Land Cover Data

For the land cover classification task, we retrieved land use/land cover data from the CORINE inventory, an harmonized classification system for Europe, launched by the European Commission (Copernicus Land Monitoring Service, 2018). In the case of Milan, approximately 94% of the building blocks can be classified into three categories: *discontinuous urban fabric*, *continuous urban fabric*, or *industrial or commercial units*. The remaining 6.1% is distributed across four minor classes, for which data are scarce. Consequently, our discussion focuses on the three predominant classes.

Within the three selected classes, *continuous urban fabric* and *discontinuous urban fabric* together account for over 96% of blocks. These two classes are defined in terms of land cover composition, relying on the proportion of artificially surfaced areas, like buildings and roads, and on the amount and type of vegetation. Continuous urban fabric is characterized by compact, densely built environments, with over 80% of the surface covered by impermeable features and scarce urban greens. In contrast, discontinuous urban fabric has lower built density (between 30% and 80%), greater built environment fragmentation, and a more significant proportion of vegetated areas and bare surfaces. Given these characteristics, the two classes provide a suitable test-bed for evaluating whether our latent representation capture geometric and spatial structure. Although our embeddings do not capture all components of the CORINE definitions (e.g., roads, vegetation), buildings' geometric properties serve as proxies for class distinctions. We claim that the ability of our embeddings to reasonably discern between the classes provides evidence that they successfully encode relevant spatial and geometric information.

We note that the remaining minority class, *industrial or commercial units*, is defined by land use rather than land cover, encompassing parcels associated with industrial, commercial, and public service functions. As a result, a wide range of land cover types can be present within this class, and its identification does not rely on consistent morphological characteristics. Therefore, urban form alone cannot reliably distinguish this category. Since our proposed embeddings are derived from blocks of building footprints, they are not expected to capture functional distinctions that are not reflected in spatial form, hence this class is discarded.

Land Cover Classification Set Up

As discussed, we consider only blocks from *continuous urban fabric* and *discontinuous urban fabric*, and train a selection of classifiers on block embeddings. Selected classifiers include Support Vector Classifier (SVC), Random Forest (RF) and Gradient Boosting (GB). We

partition our data into training and test sets following the same procedure described in Section 3.5 and generate embeddings for each partition. To perform 10-fold cross-validation, a different portion of the data was designated as test set for each fold, maintaining class balance across folds. As an initial benchmark, we adopt the model proposed by Van't Veer et al. (2019), whose depth and complexity are comparable to our approach, allowing for a fair evaluation. More advanced architectures, including transformer-based models and multimodal models (Yu et al., 2024; Li et al., 2023), will be the focus of our future investigations. While additional benchmark models will be incorporated in future work, in this paper we focus on a single baseline to establish a point of reference, while accounting for time constraints associated with this work. Van't Veer et al. (2019) design a supervised model based on elliptic Fourier descriptors (EFD) and a convolutional neural network (CNN). Polygons and multi-polygons are converted into sequences of vertices, and each vertex is concatenated with a boolean vector encoding its position in the geometry. These sequences of vertices are used to compute elliptic Fourier descriptors. For multi-polygons, this computation is performed on the largest component. A significant limitation of the Van't Veer et al. (2019) model is the inability to handle increasing geometrical complexity; particularly, polygonal geometries with interiors are automatically discarded in the original implementation. This results in a portion of our data being unusable, and therefore ignored.
We tested Van't Veer et al. (2019)'s model on different settings: classification of building blocks (definition 2), classification of building clusters (definition 3), and scale-transfer classification. The evaluation on building clusters is intended to assess the capability of the model on simpler geometric groups before applying it to more complex ones. In the scale-transfer task, the model was trained on simpler polygonal geometries (i.e., building clusters) and tested on more complex ones (i.e., building blocks): this setup resembles the conditions under which our own model is evaluated.

Results

We evaluate the performance of the models across four metrics: accuracy, balanced accuracy, F1 score, and Area Under the Curve (AUC). In particular, we report the F1 score for the *discontinuous urban fabric* class, treated as the positive class, to provide a clearer assessment of performance on the most challenging category. In the discussion, we focus on the F1 score to give equal weight to precision (i.e., the proportion of predicted positives that are actually positives) and recall (i.e., the proportion of actual positives correctly identified). The results of our experiments are reported in Table 1. Our classifiers perform comparably to or better than the benchmark model across all evaluation metrics and demonstrate significantly greater stability over the 10 folds, as indicated by the lower standard deviations. The benchmark model exhibits moderately low F1 variability on clusters (~6%), but this rises sharply to ~18%-19% on blocks, indicating that the model struggles with more complex geometries. On the contrary, our models report much more stable performances, with a F1 variability between 4.5% and 7.1%. The scale-transfer experiment shows the inability of the benchmark model to generalize across scales; conversely, our trained encoder was able to generate satisfactory representations of complex building blocks, despite being trained on simpler clusters, effectively transferring across scales. Our approach has at least two advantages allowing greater robustness:

1. By concatenating multiple polygons within the same block into a single sequence, the benchmark model breaks the topology of the building blocks. The loss of topology prevents the CNN from fully capturing spatial adjacency, continuity or fragmentation, limiting its ability to differentiate complex urban structures. In contrast, our approach preserves topological relationships, which enables improved classification performances.

2. Our encoder is a frequency-based feature extractor, hence provides global features; conversely, spatial models relying on CNNs provide local features. The capacity to generate a global representation of the block is essential for capturing contextual elements. Moreover, the differentiation between built environment characteristics in continuous and discontinuous urban fabric is primarily based on low-frequency components, i.e., broader structure, rather than high-frequency, i.e., fine geometrical details, making our encoder more suitable for the task.

Our model is designed to emphasize stable global structure through mechanisms that promote robustness to fine-scale perturbations. Hence, low-frequency components capturing geometric elements such as spatial arrangement (e.g., adjacency, continuity, broad distances), compactness, and density are well-represented in the embeddings. While fine-details representation could improve classification in specific instances (e.g., capturing irregularities in discontinuous urban fabric), the current representation is adequate for distinguishing most building-related differences between classes. Nevertheless, tasks that require precise identification of shape complexities or irregularities are likely to underperform with our embeddings due to their limited ability to capture fine-details.

Overall, our model demonstrates robustness across diverse geometrical complexities, effectively processing all available data. Moreover, the unsupervised approach enhances both efficiency and scalability by eliminating the need for manual labelling, and enables generalization at multiple scales.

It is worth noting that land cover classification is not dependent only on building configurations, and the

Table 1. Performance of classification models on the binary land cover classification task. Values represent mean ± relative standard deviation over 10-fold cross-validation. All metrics are bounded between 0 and 1, with 1 representing perfect performance. The *EFD+CNN on clusters* model represents the benchmark tested on building groups to assess its capabilities on simpler geometrical structures. The models proposed in this study are listed below the second solid horizontal line. For each metric, the highest mean value across all models is highlighted in bold.

Model	Land Cover Classification			
	Accuracy	Balanced Accuracy	F1	AUC
EFD+CNN on clusters	0.572 (±0.040)	0.573 (±0.031)	0.534 (±0.062)	0.607 (±0.038)
EFD+CNN on blocks	**0.620 (±0.102)**	**0.626 (±0.097)**	0.551 (±0.194)	0.561 (±0.154)
EFD+CNN on scale-transfer	0.460 (±0.094)	0.480 (±0.090)	0.264 (±0.182)	0.360 (±0.135)
NUFT+MLP+SVC	0.616 (±0.049)	0.617 (±0.048)	**0.602 (±0.057)**	**0.658 (±0.057)**
NUFT+MLP+RF	0.613 (±0.059)	0.611 (±0.059)	0.572 (±0.071)	0.656 (±0.060)
NUFT+MLP+GB	0.587 (±0.040)	0.587 (±0.037)	0.549 (±0.045)	0.629 (±0.053)

task has been demonstrated here purely to evaluate our unsupervised representation, using only geometric embeddings to isolate their contribution. Other urban elements, such as roads and vegetation, play a significant role in CORINE definitions, and their integration is necessary for improved performance. In this context, our embeddings can be intended as a modular component for multimodal pipelines, designed to capture building geometrical structure. Unlike handcrafted features, which require extensive domain knowledge, feature selection, and tuning, with frequent information loss, our encoder enables automated representation of urban form, preserving structural and spatial information that has been previously under-represented in multimodal models.

5 Conclusions and Future Work

In this study, we explored the potential of general-purpose representation of urban building groups and blocks. By moving beyond pre-defined spatial partitions and handcrafted features, we demonstrated the feasibility of an adapted spectral approach that learns directly from groups of building footprints. The proposed framework enables the end-to-end learning of urban form representations and produces transferable embeddings that can support multiple downstream tasks. Our results indicate that the model effectively captures global geometric patterns, including compactness and fragmentation, elongation, shape, and spatial arrangement. As shape complexity increases, the global nature of the learned representations allows this approach to recognize and cluster together coarse structural outlines across different complexity levels. The preference for global, low-frequency, geometric structure allows the model to ignore fine-scale geometric noise. This property facilitates tasks such as generalized shape characterization, but could pose challenges in more detailed-oriented tasks. However, we believe that this novel approach has great future potential. Using embeddings in this way could allow the automatic identification of: historic city areas, suburban housing, areas prone to environmental issues such as air pollution or urban heat effects, informal settlements, and so on. Without the need to manually define precise rules, this approach sees the model *learn* these patterns.

Given the exploratory nature of this study, our work primarily serves to establish the novelty and viability of this approach in urban contexts and to provide a baseline that can be used for our future work. Based on the observed limitations, we identify several ways to potentially improve the model moving forward.

- As mentioned in Section 3.5, the generalizability of our approach should be validated by applying it to cities with different urban configurations and development histories. To this end, the *one square mile* approach described in Boeing (2021) may be a useful tool for selecting diversified cities.
- Incorporating explicit frequency analysis by processing low, medium, and high frequency components in parallel could allow the model to disentangle global structure from fine-scale details.
- Replacing the MLP with architectures better suited to complex geometries, such as transformer-based models, could improve expressiveness with real-world urban data.
- Increasing the number of complex observations in the training dataset could ease the learning of intricate geometric patterns. To this end, we will evaluate including both building blocks and clusters in the training set.

Data and Software Availability

The source code is publicly available on GitHub at https://github.com/luisalopresti/BaselineBuildingRL. Building footprints are derived from OpenStreetMap, and the code to reconstruct the dataset is included in the repository. Land cover data are obtained from the CORINE inventory.

Declaration of Generative AI in writing

The authors declare that they have not used Generative AI tools in the preparation of this manuscript. All intellectual and creative work, including the analysis and interpretation of data, is original and has been conducted by the author without AI assistance.

Acknowledgements

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number 18/CRT/6049. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Atwal, K. S., Anderson, T., Pfoser, D., and Züfle, A.: Predicting building types using OpenStreetMap, Scientific Reports, 12, 19 976, https://doi.org/10.1038/s41598-022-24263-w, 2022.

Balsebre, P., Huang, W., Cong, G., and Li, Y.: City Foundation Models for Learning General Purpose Representations from OpenStreetMap, https://arxiv.org/abs/2310.00583, 2024.

Basaraner, M. and Selcuk, M.: A Structure Recognition Technique in Contextual Generalisation of Buildings and Built-up Areas, The Cartographic Journal, 45, 274–285, https://doi.org/10.1179/174327708X347773, 2008.

Bei, W., Guo, M., and Huang, Y.: A Spatial Adaptive Algorithm Framework for Building Pattern Recognition Using Graph Convolutional Networks, Sensors, 19, https://doi.org/10.3390/s19245518, 2019.

Boeing, G.: OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks, Computers, Environment and Urban Systems, Volume 65, Pages 126-139, ISSN 0198-9715, https://doi.org/10.1016/j.compenvurbsys.2017.05.004, 2017.

Boeing, G.: Spatial information and the legibility of urban form: Big data in urban morphology, International Journal of Information Management, 56, 102 013, https://doi.org/10.1016/j.ijinfomgt.2019.09.009, 2021.

Choudhury, S., Aharoni, E., Suvarna, C., Tsogsuren, I., Kreidieh, A. R., Lu, C.-T., and Arora, N.: S2Vec: Self-Supervised Geospatial Embeddings for the Built Environment, https://doi.org/10.1145/3787217, 2026.

Copernicus Land Monitoring Service: CORINE Land Cover 2018 (vector), Europe, https://doi.org/10.2909/71c95a07-e296-44fc-b22b-415f42acfdf0, 2018.

De Sabbata, S., Ballatore, A., Liu, P., and Tate, N.: Learning urban form through unsupervised graph-convolutional neural networks, in: The 2nd International Workshop on Geospatial Knowledge Graphs and GeoAI: Methods, Models, and Resources, https://doi.org/10.17605/OSF.IO/H2AWQ, 2023.

Deng, M., Tang, J., Liu, Q., and Wu, F.: Recognizing building groups for generalization: a comparative study, Cartography and Geographic Information Science, 45, 187–204, https://doi.org/10.1080/15230406.2017.1302821, 2018.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, p. 226–231, AAAI Press, 1996.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, https://doi.org/10.1109/CVPR.2016.90, 2016.

He, X., Zhang, X., and Xin, Q.: Recognition of building group patterns in topographic maps based on graph partitioning and random forest, ISPRS Journal of Photogrammetry and Remote Sensing, 136, 26–40, https://doi.org/10.1016/j.isprsjprs.2017.12.001, 2018.

Hecht, R., Meinel, G., and Buchroithner, M.: Automatic identification of building types based on topographic databases–a comparison of different data sources, International Journal of Cartography, 1, 18–31, https://doi.org/10.1080/23729333.2015.1055644, 2015.

Huang, L., Yang, Y., Zhao, X., Gao, H., and Yu, L.: Mining the Relationship between Spatial Mobility Patterns and POIs, Wireless Communications and Mobile Computing, 2018, https://doi.org/10.1155/2018/4392524, 2018.

Jiang, C. M., Lansigan, D. L. O., Marcus, P., and Nießner, M.: DDSL: Deep Differentiable Simplex Layer for Learning Geometric Signals, https://arxiv.org/abs/1901.11082, 2019.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P.: Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86, 2278–2324, https://doi.org/10.1109/5.726791, 1998.

Li, Y., Huang, W., Cong, G., Wang, H., and Wang, Z.: Urban Region Representation Learning with OpenStreetMap Building Footprints, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23, p. 1363–1373, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3580305.3599538, 2023.

Liu, T., Zhang, Z., Du, P., Wang, W., Yan, H., Qiang, B., and Xu, S.: Recognition of building group patterns using GCN and knowledge graph, Geocarto International, 40, 2436 906, https://doi.org/10.1080/10106049.2024.2436906, 2025.

Lo Presti, L. and Mooney, P.: Road network simplification to support air quality analysis, in: 33nd Annual GIS Research UK Conference (GISRUK), University of Bristol, UK, https://doi.org/10.5281/zenodo.15089554, 2025.

Mac Aodha, O., Cole, E., and Perona, P.: Presence-Only Geographical Priors for Fine-Grained Image Classification, in: Proceedings of the IEEE/cvf international conference on computer vision, pp. 9596–9606, https://doi.org/10.1109/ICCV.2019.00969, 2019.

Mai, G., Jiang, C., Sun, W., Zhu, R., Xuan, Y., Cai, L., Janowicz, K., Ermon, S., and Lao, N.: Towards general-purpose representation learning of polygonal geometries, Geoinformatica, 27, 289–340, https://doi.org/10.1007/s10707-022-00481-2, 2022.

Niu, H. and Silva, E. A.: Delineating urban functional use from points of interest data with neural network embedding: A case study in Greater London, Computers, Environment and Urban Systems, 88, 101 651, https://doi.org/10.1016/j.compenvurbsys.2021.101651, 2021.

Oostwegel, L. J., Schorlemmer, D., and Guéguen, P.: From Footprints to Functions: A Comprehensive Global and Semantic Building Footprint Dataset, Scientific Data, 12, 1699, https://doi.org/10.1038/s41597-025-06132-z, 2025.

Qin, Y., Zhao, N., Yang, J., Pan, S., Sheng, B., and Lau, R. W. H.: UrbanEvolver: Function-Aware Urban Layout Regeneration, International Journal of Computer Vision, 132, https://doi.org/10.1007/s11263-024-02030-w, 2024.

Ren, Y., Xie, Z., and Zhai, S.: Urban Land Use Classification Model Fusing Multimodal Deep Features, ISPRS International Journal of Geo-Information, 13, https://doi.org/10.3390/ijgi13110378, 2024.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography, 40, 913–929, https://doi.org/10.1111/ecog.02881, 2017.

Sun, F., Qi, J., Chang, Y., Fan, X., Karunasekera, S., and Tanin, E.: Urban Region Representation Learning with Attentive Fusion, in: 2024 IEEE 40th International Conference on Data Engineering (ICDE), pp. 4409–4421, https://doi.org/10.1109/ICDE60146.2024.00336, 2024.

Taubenböck, H., Debray, H., Qiu, C., Schmitt, M., Wang, Y., and Zhu, X.: Seven city types representing morphologic configurations of cities across the globe, Cities, 105, 102 814, https://doi.org/10.1016/j.cities.2020.102814, 2020.

United Nations: AI for Spatial Mapping and Analysis: GeoAI Toolkit for Urban Planners, available at: https://unitac.un.org/en/media/687, last access: 25 March 2026, 2026.

United Nations, Department of Economic and Social Affairs: World Urbanization Prospects 2025, available at: https://population.un.org, last access: 15 December 2025, 2025.

Van den Oord, A., Li, Y., and Vinyals, O.: Representation Learning with Contrastive Predictive Coding, https://arxiv.org/abs/1807.03748, 2019.

Van der Maaten, L. and Hinton, G.: Visualizing Data using t-SNE, Journal of Machine Learning Research, 9, 2579–2605, http://jmlr.org/papers/v9/vandermaaten08a.html, 2008.

Van't Veer, R., Bloem, P., and Folmer, E.: Deep Learning for Classification Tasks on Geospatial Vector Polygons, https://arxiv.org/abs/1806.03857, 2019.

Wang, X., Cheng, T., Law, S., Zeng, Z., Yin, L., and Liu, J.: Multi-modal contrastive learning of urban space representations from POI data, Computers, Environment and Urban Systems, 120, 102 299, https://doi.org/10.1016/j.compenvurbsys.2025.102299, 2025.

Yan, X., Ai, T., Yang, M., and Yin, H.: A graph convolutional neural network for classification of building patterns using spatial vector data, ISPRS Journal of Photogrammetry and Remote Sensing, 150, 259–273, https://doi.org/10.1016/j.isprsjprs.2019.02.010, 2019.

Yan, X., Ai, T., Yang, M., and Tong, X.: Graph convolutional autoencoder model for the shape coding and cognition of buildings in maps, International Journal of Geographical Information Science, 35, 490–512, https://doi.org/10.1080/13658816.2020.1768260, 2021.

Yu, D., Hu, Y., Li, Y., and Zhao, L.: PolygonGNN: Representation Learning for Polygonal Geometries with Heterogeneous Visibility Graph, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24, p. 4012–4022, ACM, https://doi.org/10.1145/3637528.3671738, 2024.

Zhang, Y., Huang, W., Yao, Y., Gao, S., Cui, L., and Yan, Z.: Urban region representation learning with human trajectories: a multi-view approach incorporating transition, spatial, and temporal perspectives, GIScience & Remote Sensing, 61, 2387 392, https://doi.org/10.1080/15481603.2024.2387392, 2024.