ISPRS Journal of Photogrammetry and Remote Sensing 203 (2023) 135–164

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Screening the stones of Venice: Mapping social perceptions of cultural significance through graph-based semi-supervised classification

Nan Bai [a,*], Pirouz Nourian [b], Renqian Luo [c], Tao Cheng [d], Ana Pereira Roders [a]

[a] UNESCO Chair in Heritage and Values: Heritage and the Reshaping of Urban Conservation for Sustainability, Department of Architectural Engineering and Technology, Delft University of Technology, 2628BL Delft, The Netherlands
[b] Faculty of Geo-Information Science and Earth Observation, University of Twente, 7522NH Enschede, The Netherlands
[c] AI4Science, Microsoft Research, 100080 Beijing, China
[d] SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineer, University College London, Gower Street, WC1E 6BT London, The United Kingdom

ARTICLE INFO

Keywords:
Social media data
Multi-modal machine learning
Graph Neural Networks
Spectral centrality
Heritage values and attributes
Label diffusion

ABSTRACT

Mapping cultural significance of heritage properties in urban environment from the perspective of the public has become an increasingly relevant process, as highlighted by the 2011 UNESCO Recommendation on the Historic Urban Landscape (HUL). With the ubiquitous use of social media and the prosperous developments in machine and deep learning, it has become feasible to collect and process massive amounts of information produced by online communities about their perceptions of heritage as social constructs. Moreover, such information is usually inter-connected and embedded within specific socioeconomic and spatiotemporal contexts. This paper presents a methodological workflow for using semi-supervised learning with graph neural networks (GNN) to classify, summarize, and map cultural significance categories based on user-generated content on social media. Several GNN models were trained as an ensemble to incorporate the multi-modal (visual and textual) features and the contextual (temporal, spatial, and social) connections of social media data in an attributed multi-graph structure. The classification results with different models were aligned and evaluated with the prediction confidence and agreement. Furthermore, message diffusion methods on graphs were proposed to aggregate the post labels onto their adjacent spatial nodes, which helps to map the cultural significance categories in their geographical contexts. The workflow is tested on data gathered from Venice as a case study, demonstrating the generation of social perception maps for this UNESCO World Heritage property. This research framework could also be applied in other cities worldwide, contributing to more socially inclusive heritage management processes. Furthermore, the proposed methodology holds the potential of diffusing any human-generated location-based information onto spatial networks and temporal timelines, which could be beneficial for measuring the safety, vitality, and/or popularity of urban spaces.

1. Introduction

Documenting and mapping the values (cultural significance) of cities have always been an important task in the practice of urban conservation (Zancheti and Jokilehto, 1997; ICOMOS, 2013). As an art critic, historian, writer, polymath, and a pioneer in heritage conservation, John Ruskin openly expressed and actively promoted the cultural significance of the grandiose architecture on the Venetian island in his three-volume masterpiece *The Stones of Venice* (Ruskin, 1879; Ruskin and Quill, 2015). Through several visits to Venice, Ruskin was attracted by the buildings, monuments, sculptures, and building elements, especially those dating from the era of Byzantine and Gothic. In fear of losing its cultural significance by industrial modernization and destructive restorations, Ruskin tirelessly documented every stone of Venice with his detailed drawings and enthusiastic guide for the readers on what to appreciate and value in future visits. However, the expressions Ruskin used can be subjective and reflect his personal tastes, which is evident in his objection against the "colourless" Renaissance buildings. Like all other visitors, the words of Ruskin describing Venice were regarded as a myth, a fiction, and a symbolic landscape, reflecting his own imagination of this idealized city (Cosgrove, 1982; Psarra, 2018). Turning the argument around, like Ruskin, all the other visitors and residents in Venice are also qualified to express the values the city conveys to them. Psarra (2018) argues that "[*a*]*ny effort to describe Venice runs the risk of confusing the city with the words and the images that describe it*", bringing up another question about what these "words and images" really are about.

* Corresponding author.
E-mail address: n.bai@tudelft.nl (N. Bai).

https://doi.org/10.1016/j.isprsjprs.2023.07.018
Received 22 December 2022; Received in revised form 21 June 2023; Accepted 18 July 2023
Available online 4 August 2023

The modern era of Social Media has given more opportunities and challenges to the process of collecting and mapping cultural significance from the perspective of general public. This is because social media has made possible the open publication of ideas, opinions, and emotions by everyone among the online communities with their own "words and images" (Cartwright, 2010). Like the pieces of stones observed by Ruskin, those posts on social media could be understood as "digital notes of stones" to be screened and inspected to dig valuable messages. Analysing such massive data can help collect information on the cultural significance (i.e., the values of cultural heritage embodied in the places for all generations) conveyed to the general public, map knowledge from alternative perspectives other than the expert-based authorized heritage discourse, and construct an inclusive heritage management plan respecting the collective opinions (ICOMOS, 2013; Aggarwal, 2011; Amato et al., 2016; Bigne et al., 2021; Bai et al., 2021b). This aligns well with the goals and objectives set by the 2011 UNESCO Recommendation on the Historic Urban Landscape (HUL) (UNESCO, 2011; Bandarin and Van Oers, 2012; Pereira Roders, 2019). Among all the information and knowledge to be extracted and mapped, heritage values (why to conserve) and heritage attributes (what to conserve) are arguably the most informative ones to fully understand the cultural significance of a heritage property, being listed or not, e.g., see Pereira Roders (2007), Tarrafa Silva and Pereira Roders (2010) and Veldpaus (2015). Ginzarly et al. (2019) demonstrates an example in this line to map the HUL values revealed on Flickr by manually checking the post contents. In the past decades, the advances in Machine Learning (ML) and Deep Learning (DL), especially Multimodal Machine Learning focusing on fusing information from different modalities (such as texts and images), have enabled similar analyses at larger scales (LeCun et al., 2015; Baltrusaitis et al., 2019; Cao et al., 2020). In order to extract and map the most representative categories of descriptions and/or images of a place, earlier studies constructed textual and visual information from social media posts with handcrafted or learned features (Crandall et al., 2009; Monteiro et al., 2014; Huang and Li, 2016; Lai et al., 2017; Boy and Uitermark, 2017), while recent studies have been updating the process with neural network models pre-trained on generic tasks for generalizable results (Gomez et al., 2019; Zhang et al., 2019b; Kang et al., 2021; Cho et al., 2022; Bai et al., 2022; Zhang et al., 2022b; Wang et al., 2022a; He et al., 2022).

However, two challenges remain for the approach of mapping cultural significance to be broadly applied in heritage and urban studies: (1) the raw user-generated data collected from social media are usually hard to annotate especially when the labels need complex expert knowledge; (2) the time-stamped and geo-tagged posts are usually scattered in space, which need to be further aggregated and summarized into higher-level spatial units, resulting in maps that are comprehensible by planners and decision-makers. Since social media posts are embedded in socioeconomic and spatiotemporal contexts (i.e., in explicit or intrinsic graph structures denoting the connections of posts such as located in nearby places, posted in consecutive time periods, and owned by similar social groups), both challenges can be handled with the emerging fields of Semi-supervised Machine Learning on Graphs with Graph Neural Network (GNN) (Zhang and Cheng, 2020; Ma and Tang, 2021; Wu et al., 2022; Xu et al., 2022). Different from conventional supervised learning, semi-supervised learning models also have access to features from unlabelled data during training process without knowing their "true" labels (Zhou and Li, 2010). This is proved to be effective especially on graphs since neighbours on graphs are assumed to be similar both in the feature space and the label space (Zhu and Ghahramani, 2002; Kipf and Welling, 2016; Xu et al., 2022). With spatial data in physical space, such similarity is expressed as the rule of the *First Law of Geography* (Tobler, 1970), that nearby things are generally similar to, and therefore, more likely to influence each other.

This paper aims to explore the use of graph-based semi-supervised classification to spatially map the cultural significance categories of cities with multi-modal social media data embedded in a graph structure. To reach the aim, three research questions are explored, becoming the three main components of the workflow proposed in this paper:

1. How can graph-based semi-supervised classification help to classify a partially labelled multi-modal social media dataset concerning location-based categorical information in a city?
2. How can an ensemble of trained models help to further improve classification performance and reliability?
3. How can the labels assigned for the posts be aggregated onto the spatial network of a city in order to map the categorical information (the perceived cultural significance)?

The scope and the approach of this study are highly related to Liu and De Sabbata (2021), where the authors presented a framework for using GNN to classify multi-modal features into user-defined label sets. Whereas Liu and De Sabbata (2021) focused on exploring the effects of different graph construction methods for only one specific type of GNN model (i.e., Graph Convolutional Network) and the mapping procedure was only a showcase of randomly sampled scatter points without further spatial aggregation and applicational analyses, this study has the following further contributions:

- A few Deep Learning models are trained on a semi-supervised classification task about cultural significance with partially labelled multi-modal graph-based datasets, and the soft-label predictions of individual models were aggregated into ensemble results, keeping track of the confidence and agreement of the models, as a measure of reliability;
- The obtained post labels are further aggregated into spatial nodes and diffused on a spatial network based on the geographical/topological proximity, effectively summarizing the information into a set of spatial maps for cultural significance categories;
- Detailed analyses on the spatial and aspatial distributions of the cultural significance categories, as well as the association of input features and output categories are provided, informative for future inclusive heritage management processes.

The workflow demonstrated in this paper with the special case of heritage cultural significance can be easily generalized in other use cases for spatially diffusing and mapping any human-generated features and labels, which can be extended to the evaluation of spatial safety, vitality, and/or architectural style in urban spaces (Cheng and Wicks, 2014; Zhang et al., 2022a; Sun et al., 2022).

2. Methodology

2.1. Case study

To relate to the metaphor of the title and its relationship with Ruskin's controversial masterpiece *The Stones of Venice* (Ruskin, 1879; Ruskin and Quill, 2015), this study selects Venice as a case study to test the methodological framework. *Venice and its Lagoon* was inscribed in the UNESCO World Heritage List in 1987 fulfilling all first six selection criteria of Outstanding Universal Value (OUV) related to cultural heritage (UNESCO, 1972, 2008; Jokilehto, 2007). Despite its status as a cultural heritage property, its special urban typology and intimate relationship with the water give the city strong clues of natural values (Bai et al., 2022), making it a popular tourism destination of diverse interests, which also means that it may suffer from the mass-tourism (Urry and Larsen, 2011; Bertocchi and Visentin, 2019). Meanwhile, Venice can be found in various academic publications and non-academic fictions, as well as voluntary comments on social media platforms, providing abundant information from all sorts of perspectives (Calvino, 1978; Cosgrove, 1982; Bigne et al., 2021). The city itself is also a product of top-down conscious city planning (statecraft) and bottom-up collective community building (city-craft) (Psarra,

Table 1
The distribution of cultural significance categories as OUV selection criteria and heritage attributes in the training sets.

Dataset	*VEN*	*VEN-XL*		*VEN*	*VEN-XL*
OUV Selection Criteria (within top-3 entries)	(361)	(11,569)	**Heritage attributes** (within top-1 entries)	(361)	(11,569)
Criterion (i) - Masterpiece	172 (15.9%)	2463 (7.1%)	Monument and buildings	69 (19.1%)	1507 (13.0%)
Criterion (ii) - Influence	188 (17.4%)	4704 (13.6%)	Building elements	71 (19.7%)	1501 (13.0%)
Criterion (iii) - Testimony	247 (22.8%)	9864 (28.4%)	Urban form elements	101 (28.0%)	2636 (22.8%)
Criterion (iv) - Typology	261 (24.1%)	8578 (24.7%)	Urban scenery	6 (1.7%)	113 (1.0%)
Criterion (v) - Land-use	7 (0.6%)	54 (0.2%)	Natural features and		
Criterion (vi) - Association	205 (18.9%)	8921 (25.7%)	Landscape scenery	30 (8.3%)	2051 (17.7%)
Criterion (vii) - Natural Beauty	1 (0.1%)	58 (0.2%)	Interior scenery	25 (6.9%)	480 (4.1%)
Criterion (viii) - Geological Process	0 (0.0%)	18 (0.1%)	People's activity and		
Criterion (ix) - Ecological Process	1 (0.1%)	19 (0.1%)	Association	49 (13.6%)	2457 (21.2%)
Criterion (x) - Bio-diversity	1 (0.1%)	28 (0.1%)	Gastronomy	9 (2.5%)	139 (1.2%)
Others - Not related	0 (0.0%)	0 (0.0%)	Artifact products	1 (0.3%)	685 (5.9%)

Table 2
Descriptive overview of the data used for this study previously collected by Bai et al. (2022).

Dataset	*VEN*			*VEN-XL*		
	Number/Count	Rate/Proportion		Number/Count	Rate/Proportion	
Nodes	2 951	–		80,963	–	
Nodes with Visual Features	2 951	100%		80,963	100%	
Nodes with Textual Features	1 761	59.7%		49,823	61.5%	
Nodes with OUV Labels	756	25.6%		25,771	31.8%	
Nodes with Heritage Attribute Labels	1 356	45.9%		37,289	46.1%	
Nodes with Both Types of Labels	361	12.2%		11,569	14.3%	
	Number/Count	Average degree	Density	Number/Count	Average degree	Density
Temporal Links	249,120	84.4	.057	35,527,354	438.8	.011
Social Links	242,576	82.2	.056	38,170,651	471.5	.012
Spatial Links	221,414	75.0	.051	101,046,098	1248.1	.031
Simple Composed Links[a]	534,513	181.1	.123	145,005,270	1791.0	.044

[a]Multiple links among two nodes leads to only one link in the simple composed graph.

2018), both firmly embedded in a spatiotemporal and socioeconomic context. All these characteristics make Venice a representative case study to demonstrate the utility of the proposed framework. Yet, it is also important to notice that the selection of Venice as the case study is only a pragmatic choice, and hypothetically the framework should be generalizable in other cities containing World Heritage, similar to Psarra's argument, that Venice could be considered as a prototype of other global cities (Psarra, 2018).

2.2. Data

This study uses the open datasets *Heri-Graphs-Venice (VEN)* and *Venice-Large (VEN-XL)* introduced by Bai et al. (2022), where multi-modal information from the social media platform Flickr is collected, containing visual and textual features, temporal, social, and spatial contexts (as a multi-graph), as well as partially-labelled pseudo-labels for cultural significance categories based on model confidence. In their definition, cultural significance was specified with two concepts as soft labels, effectively providing two probability distribution vectors: an 11-class OUV selection criteria (referred to from here on as OUV for simplicity) category (UNESCO, 1972, 2008; Jokilehto, 2008; Bai et al., 2021a), and a 9-class heritage attributes (HA) category (Veldpaus, 2015; Gustcoven, 2016; Ginzarly et al., 2019), both listed in Table 1. Since Flickr is an image-sharing platform and textual information is not mandatory during posting, both datasets collected therefrom were better equipped with visual features as 982-dimensional stacked vectors of a few pre-trained model outputs, and only about half of data samples contained valid BERT-based textual features as 771-dimensional vectors.

Within the two datasets, the lite version *VEN* was already formatted as a multi-graph with three types of undirected weighted links (temporal, social, and spatial) showing the contextual connections among the nodes representing posts on Flickr. However, the larger version *VEN-XL* was only provided with the nodal features because of the large memory requirement to construct adjacency matrices with a huge number of nodes. Following the guidelines given by Bai et al. (2022), this paper also constructed multi-graph mini-batches for *VEN-XL* in Pytorch-Geometric library (Fey and Lenssen, 2019) using sparse matrices as graph structure (Yuster and Zwick, 2005). An overview of both datasets is given in Table 2. The label rates (.122/.143) of the datasets are comparable with common semi-supervised learning datasets in graph neural networks such as Citeer (.036) and Cora (.052) (Kipf and Welling, 2016; Yang et al., 2016). Note *VEN-XL* has a larger average degree for nodes with all types of links, yet the multi-graphs are less dense than the lite *VEN* dataset.

As a summary, the datasets in this study have three challenges for the semi-supervised classification task: (1) only partial labels are available for the categories of interest, requiring the unlabelled nodes to be tagged; (2) only partial features are available for some nodes, requiring the models to learn as much as possible from their neighbours on graphs; (3) the VEN-XL dataset is too large to conduct training and inference directly, requiring sampling of subgraphs. All these characteristics of the datasets entail that both transductive (training and inference on the same graph) and inductive (inference on unseen [sub-] graphs) semi-supervised learning on graphs (Yang et al., 2016; Liu and De Sabbata, 2021) are indispensable, reflecting the scope and necessity of this study. For both datasets, the nodes with both types of labels (OUV and HA) are treated as the training sets (361 for *VEN*; 11,569 for *VEN-XL*), and the nodes with only one type of labels are randomly and evenly separated as validation sets (695; 19,961) and test sets (695; 19,961), while the remainder of the nodes is considered as unlabelled data (1200; 29,472). In the training sets, all essential categories are present, though the distribution is unbalanced, as presented in Table 1.

2.3. Problem definition

The workflow proposed in this paper is visualized in Fig. 1. The input data from two databases *VEN* and *VEN-XL* are: (1) a partially-labelled attributed multi-graph about the inter-related social media

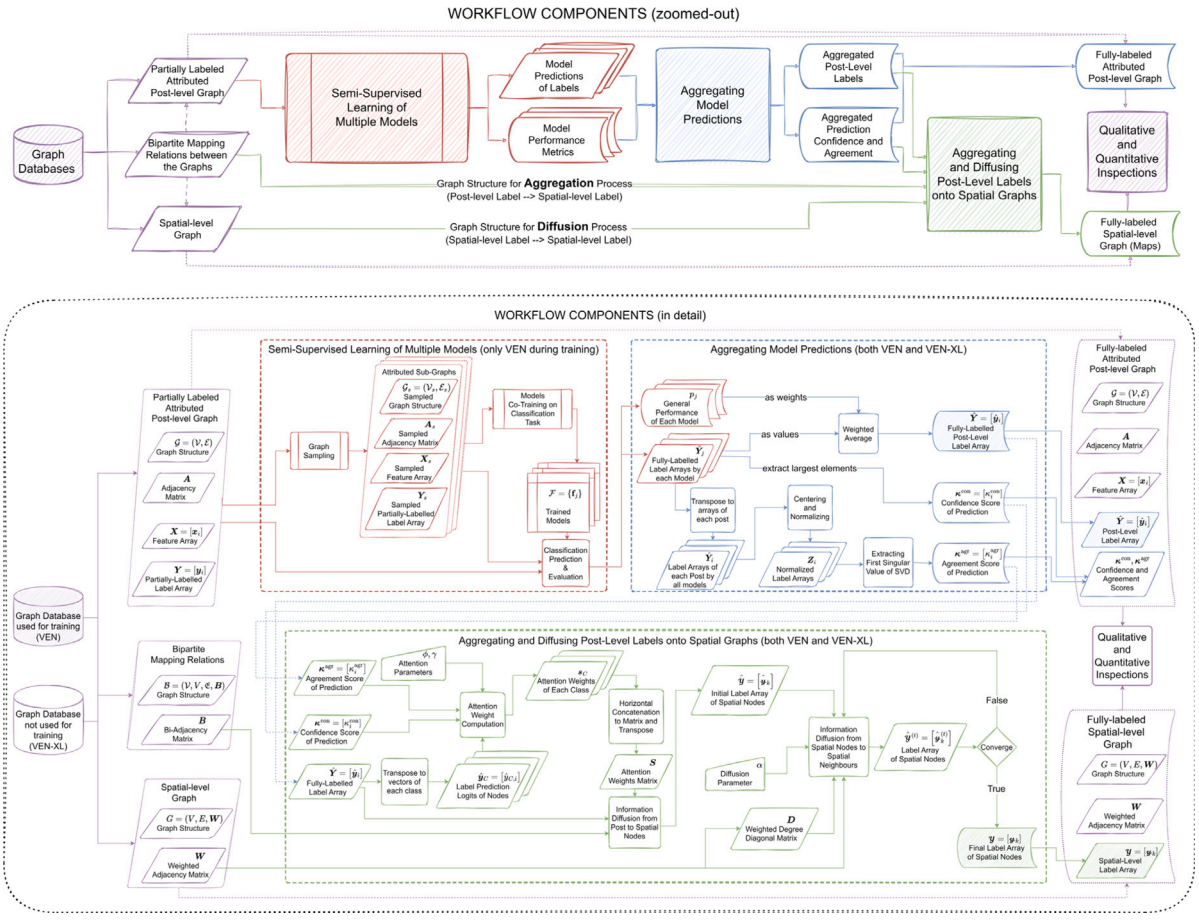

Fig. 1. The general methodological workflow proposed in this paper, both as zoomed-out high-level modulated framework in the upper part, and as a detailed workflow with mathematical notations in the lower part to be instantiated in the texts. Only the lite dataset *VEN* is used to train the models in the first step of semi-supervised learning, while the large dataset *VEN-XL* is directly used for inference and later steps. The indices i, j, k are respectively a generic example of the posts $v_i \in \mathcal{V}$, the models $\mathbf{f}_j \in \mathcal{F}$, and the spatial intersection nodes $v_k \in V$.

posts; (2) an assignment bipartite graph with relations mapping the posts to their closest street intersections (spatial nodes); (3) a topological representation of the spatial network as a weighted undirected graph marking the proximity of the street intersections. After three main components, i.e., (1) semi-supervised learning of multiple models co-trained in a classification task (Section 2.3.2), (2) aggregating the prediction outputs as soft labels of those models (Section 2.3.3), and (3) aggregating and diffusing the post-level labels on the spatial graph (Section 2.3.4), two outputs are obtained: (1) a graph fully-labelled on all post-level nodes together with confidence and agreement scores based on model performance; (2) a graph fully-labelled on spatial-level nodes summarizing the information of nearby posts and proximate spatial neighbours. Both outcomes are tested with qualitative and quantitative inspections (Section 3.3). The graph structures are conceptually visualized in Fig. 3. The process will be formally described in the following Sections. The relevant works concerning the proposed workflow will be discussed in Section 5.4.

2.3.1. General notations of attributed graphs

Since the data structure is exactly the same for *VEN* and *VEN-XL* except for the sample size, this section will describe the general notation system eligible for both datasets. For each dataset, an undirected multi-graph $\mathcal{G} = (\mathcal{V}, \{\mathcal{E}^{\text{TEM}}, \mathcal{E}^{\text{SPA}}, \mathcal{E}^{\text{SOC}}\})$ with three types of links (temporal, spatial, and social, as mentioned in Section 2.2) represents its contextual structure, where $\mathcal{V} = \{v_i\}, i \in [0, K)$ is the node set of all the posts collected and K is the total number of posts, and $(v_i, v_{i'}) \in \mathcal{E}^{(*)} \subseteq \mathcal{V} \times \mathcal{V}, \mathcal{E}^{(*)} \in \{\mathcal{E}^{\text{TEM}}, \mathcal{E}^{\text{SPA}}, \mathcal{E}^{\text{SOC}}\}$ is a link marking one type of contextual relations among the posts. For simplicity, the link weights in Bai et al. (2022) are omitted, resulting in binary adjacency matrices $\boldsymbol{A}^{(*)} := [A^{(*)}_{i,i'}] \in \{0,1\}^{K\times K}, \boldsymbol{A}^{(*)} \in \{\boldsymbol{A}^{\text{TEM}}, \boldsymbol{A}^{\text{SPA}}, \boldsymbol{A}^{\text{SOC}}\}$, where all the links $(v_i, v_{i'})$ with an original weight larger than 0 will lead to $A^{(*)}_{i,i'} = 1$, otherwise $A^{(*)}_{i,i'} = 0$. Moreover, a simple composed graph $\mathcal{G}' = (\mathcal{V}, \mathcal{E})$ could be obtained by merging the adjacency matrices into $\boldsymbol{A}$, so that $\boldsymbol{A} = (\boldsymbol{A}^{\text{TEM}} > 0) \bigvee (\boldsymbol{A}^{\text{SPA}} > 0) \bigvee (\boldsymbol{A}^{\text{SOC}} > 0) \in \{0,1\}^{K\times K}$. In this simple composed graph $\mathcal{G}'$, a link would exist if at least one contextual type of links exists between two nodes in the multi-graph $\mathcal{G}$.

For all the nodes in the graph $\mathcal{G}$, a 2D feature array $\boldsymbol{X} := [\boldsymbol{x}_i]_{i\in[0,K)} = \begin{bmatrix}\boldsymbol{X}^{\text{vis}} \\ \boldsymbol{X}^{\text{tex}}\end{bmatrix} \in \mathbb{R}^{1753\times K}$ would exist, where $\boldsymbol{x}_i \in \mathbb{R}^{1753\times 1}$ is a vector representing the features of node v_i, $\boldsymbol{X}^{\text{vis}} \in \mathbb{R}^{982\times K}, \boldsymbol{X}^{\text{tex}} \in \mathbb{R}^{771\times K}$ are respectively the visual and textual features, and $\begin{bmatrix}\cdot \\ \cdot\end{bmatrix}$ is the vertical concatenation operation of arrays. In cases where no textual data was available for a post node, the corresponding entries in vector $\boldsymbol{x}_i$ would be all zeros, dividing the nodes $\mathcal{V}$ into two sub-clusters $\mathcal{V}_{\text{tex}+}, \mathcal{V}_{\text{tex}-} \subset \mathcal{V}$, with or without textual data.

Since pseudo-labels for posts were respectively provided for a different subset of $\mathcal{V}$ concerning OUV and HA, four sub-clusters $\mathcal{V}_{\text{V}+,\text{A}+}$, $\mathcal{V}_{\text{V}+,\text{A}-}, \mathcal{V}_{\text{V}-,\text{A}+}, \mathcal{V}_{\text{V}-,\text{A}-} \subset \mathcal{V}$ could be categorized, as they have different label arrays:

- For nodes with both labels in $\mathcal{V}_{\text{V}+,\text{A}+}$, the label array would be $\boldsymbol{Y}_{\text{V}+,\text{A}+} = \begin{bmatrix}\boldsymbol{y}_i^{\text{OUV}} \\ \boldsymbol{y}_i^{\text{HA}}\end{bmatrix}_{v_i\in\mathcal{V}_{\text{V}+,\text{A}+}}$, where $\boldsymbol{y}_i^{\text{OUV}} \in [0,1]^{11\times 1}, \boldsymbol{y}_i^{\text{HA}} \in$

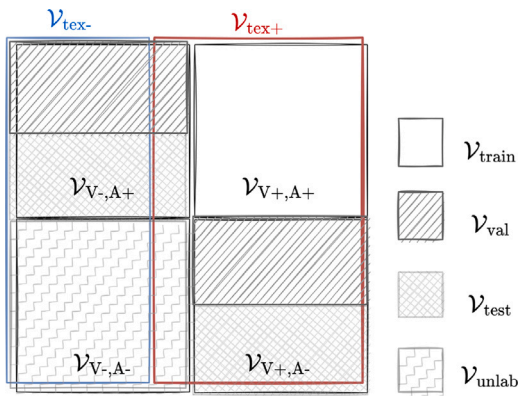

Fig. 2. The Venn Diagram showing the logic relations of the three types of sub-clustering of nodes in $\mathcal{V}$. The relationship described in Eqs. (1) and (2) are visualized.

$[0,1]^{9\times1}$ are respectively a column-stochastic vector denoting the soft labels of node v_i for OUV and HA categories;

- For nodes with only OUV labels in $\mathcal{V}_{\text{V+,A-}}$, the label array would be $\boldsymbol{Y}_{\text{V+,A-}} = \left[\boldsymbol{y}_i^{\text{OUV}}\right]_{v_i\in\mathcal{V}_{\text{V+,A-}}}$;
- For nodes with only HA labels in $\mathcal{V}_{\text{V-,A+}}$, the label array would be $\boldsymbol{Y}_{\text{V-,A+}} = \left[\boldsymbol{y}_i^{\text{HA}}\right]_{v_i\in\mathcal{V}_{\text{V-,A+}}}$;
- For nodes with in $\mathcal{V}_{\text{V-,A-}}$, there is no label array.

Note the following relationship holds for the sub-clusters:

$$\begin{aligned}&(\mathcal{V}_{\text{V+,A+}}\cup\mathcal{V}_{\text{V+,A-}})\subset\mathcal{V}_{\text{tex+}},\\&(\mathcal{V}_{\text{V-,A+}}\cup\mathcal{V}_{\text{V-,A-}})\cap\mathcal{V}_{\text{tex+}}\neq\emptyset,\\&(\mathcal{V}_{\text{V-,A+}}\cup\mathcal{V}_{\text{V-,A-}})\cap\mathcal{V}_{\text{tex-}}\neq\emptyset,\end{aligned}\tag{1}$$

meaning that having textual features as input is a necessary but not sufficient condition of having the OUV label.

2.3.2. Semi-supervised training on sampled graphs

As described in Section 2.2, the nodes in $\mathcal{V}$ are further split into training set $\mathcal{V}_{\text{train}}$, validation set $\mathcal{V}_{\text{val}}$, test set $\mathcal{V}_{\text{test}}$, and unlabelled set $\mathcal{V}_{\text{unlab}}$, where:

$$\begin{aligned}&\mathcal{V}_{\text{train}}=\mathcal{V}_{\text{V+,A+}},\\&\mathcal{V}_{\text{unlab}}=\mathcal{V}_{\text{V-,A-}},\\&\mathcal{V}_{\text{val}}\cup\mathcal{V}_{\text{test}}=\mathcal{V}_{\text{V+,A-}}\cup\mathcal{V}_{\text{V-,A+}},\\&|\mathcal{V}_{\text{val}}|=|\mathcal{V}_{\text{test}}|.\end{aligned}\tag{2}$$

The semi-supervised learning task in this paper is to use the training nodes $\mathcal{V}_{\text{train}}$ and teach a group of models to learn the mapping functions within a candidate model set $\mathcal{F}=\{\mathbf{f}_j\}, j\in[0,|\mathcal{F}|)$ from input features $\boldsymbol{X}$ to output labels $\boldsymbol{Y}$, tune the hyper-parameters and select the optimal models based on their performance on the validation nodes $\mathcal{V}_{\text{val}}$, evaluate the generalizability of the models on unseen test data on $\mathcal{V}_{\text{test}}$, and apply the trained models to generate predicted soft labels $\hat{\boldsymbol{Y}}=\left[\hat{\boldsymbol{y}}_i\right]_{v_i\in\mathcal{V}}$ for all nodal data including the ones in $\mathcal{V}_{\text{unlab}}$. The logic relations among the three types of clustering of the node set $\mathcal{V}$ mentioned in Eqs. (1) and (2) are illustrated in the Venn Diagram of Fig. 2.

For both efficiency and generalizability, sub-graphs are strategically sampled from the original graphs to train the models: $\mathcal{G}_s=(\mathcal{V}_s,\{\mathcal{E}_s^{\text{TEM}},\mathcal{E}_s^{\text{SPA}},\mathcal{E}_s^{\text{SOC}}\})$ or $\mathcal{G}_s=(\mathcal{V}_s,\mathcal{E}_s)$ with respectively sampled adjacency matrices $\boldsymbol{A}_s^{(*)},\boldsymbol{A}_s$ and feature array $\boldsymbol{X}_s$, where $\mathcal{V}_s\subseteq\mathcal{V},\mathcal{E}_s\subseteq\mathcal{E},\mathcal{E}_s^{(*)}\subseteq\mathcal{E}^{(*)}$, depending on if the models would use the multi-graph structure or the simple composed one. For each training epoch, non-repetitive mini-batches of nodes $\mathcal{V}_{\text{batch}}\subset\mathcal{V}_s$ are used as base nodes to sample several different sub-graphs $\mathcal{G}_s$. Then the training loss $\mathcal{L}_{\text{train}}$ of any model $\mathbf{f}_j$ with model parameter $\boldsymbol{\Theta}_j$ for each mini-batch $\mathcal{V}_{\text{batch}}$ could be described as:

$$\mathcal{L}_{\text{train}}(\boldsymbol{\Theta}_j,\mathcal{V}_{\text{batch}})=\sum_{v_i\in\mathcal{V}_{\text{batch}}\cap\mathcal{V}_{\text{train}}}\left(\ell(\hat{\boldsymbol{y}}_{j,i}^{\text{OUV}},\boldsymbol{y}_i^{\text{OUV}})+\omega_{\text{V/A}}\ell(\hat{\boldsymbol{y}}_{j,i}^{\text{HA}},\boldsymbol{y}_i^{\text{HA}})\right),\tag{3}$$

$$\hat{\boldsymbol{y}}_{j,i}:=\begin{bmatrix}\hat{\boldsymbol{y}}_{j,i}^{\text{OUV}}\\\hat{\boldsymbol{y}}_{j,i}^{\text{HA}}\end{bmatrix}=\begin{bmatrix}\text{softmax}(\boldsymbol{z}_{j,i}^{\text{OUV}})\\\text{softmax}(\boldsymbol{z}_{j,i}^{\text{HA}})\end{bmatrix},\tag{4}$$

$$\mathbf{1}_{11\times1}^{\text{T}}\hat{\boldsymbol{y}}_{j,i}^{\text{OUV}}=\mathbf{1}_{9\times1}^{\text{T}}\hat{\boldsymbol{y}}_{j,i}^{\text{HA}}=1,\tag{5}$$

$$\text{and }\boldsymbol{z}_{j,i}:=\begin{bmatrix}\boldsymbol{z}_{j,i}^{\text{HV}}\\\boldsymbol{z}_{j,i}^{\text{HA}}\end{bmatrix}=\mathbf{f}_j(\boldsymbol{A}_s,\boldsymbol{X}_s;\boldsymbol{\Theta}_j)_i,\tag{6}$$

where ℓ is a loss function comparing the similarity of two vectors, such as cross-entropy (Rubinstein and Kroese, 2013), $\omega_{\text{V/A}}$ is a scalar parameter balancing the importance of OUV and HA categories during training, $\hat{\boldsymbol{y}}_{j,i}^{\text{OUV}}\in[0,1]^{11\times1},\hat{\boldsymbol{y}}_{j,i}^{\text{HA}}\in[0,1]^{9\times1}$ are respectively predicted stochastic label vectors for OUV and HA by the j_{th} model on the i_{th} example, and $\boldsymbol{z}_{j,i}^{\text{OUV}}\in\mathbb{R}^{11\times1},\boldsymbol{z}_{j,i}^{\text{HA}}\in\mathbb{R}^{9\times1}$ are respectively components of the model output vector $\boldsymbol{z}_{j,i}\in\mathbb{R}^{20\times1}$. Notice that the two objectives of classifying OUV and HA are trained together with a shared model architecture and are only distinguished before final loss computation, instead of having two separate models. This is assumed to be more generalizable and could capture more information on the associations between the two closely-related topics.

While evaluating the model performance on validation set $\mathcal{V}_{\text{val}}$ (and eventually on test set $\mathcal{V}_{\text{test}}$), the computation of the scores $\mathcal{L}_{\text{val}}^{\text{OUV}}$ and $\mathcal{L}_{\text{val}}^{\text{HA}}$ respectively on OUV and HA categories would be further distinguished as:

$$\mathcal{L}_{\text{val}}^{\text{OUV}}(\boldsymbol{\Theta}_j)=\frac{\sum_{\mathcal{V}_{\text{batch}}\subset\mathcal{V}_{\text{val}}}\sum_{v_i\in\mathcal{V}_{\text{batch}}\cap\mathcal{V}_{\text{V+,A-}}}\ell_{\text{V}}(\hat{\boldsymbol{y}}_{j,i}^{\text{OUV}},\boldsymbol{y}_i^{\text{OUV}})}{\left|\mathcal{V}_{\text{val}}\cap\mathcal{V}_{\text{V+,A-}}\right|},\tag{7}$$

$$\mathcal{L}_{\text{val}}^{\text{HA}}(\boldsymbol{\Theta}_j)=\frac{\sum_{\mathcal{V}_{\text{batch}}\subset\mathcal{V}_{\text{val}}}\sum_{v_i\in\mathcal{V}_{\text{batch}}\cap\mathcal{V}_{\text{V-,A+}}}\ell_{\text{A}}(\hat{\boldsymbol{y}}_{j,i}^{\text{HA}},\boldsymbol{y}_i^{\text{HA}})}{\left|\mathcal{V}_{\text{val}}\cap\mathcal{V}_{\text{V-,A+}}\right|},\tag{8}$$

where ℓ_{V} and ℓ_{A} are topic-specific evaluation metrics for both classification tasks which will be introduced in Section 3.2.2. For each batch $\mathcal{V}_{\text{batch}}\subset\mathcal{V}_{\text{val}}$, a new sample sub-graph $\mathcal{G}_s$ is used to compute the soft labels $\hat{\boldsymbol{y}}_{j,i}^{\text{OUV}},\hat{\boldsymbol{y}}_{j,i}^{\text{HA}}$.

2.3.3. Aggregating prediction outputs

Assume the semi-supervised learning process mentioned in Section 2.3.2 trains all models in $\mathcal{F}=\{\mathbf{f}_j\}$ properly and they generate a set of well-fit label arrays $\left\{\hat{\boldsymbol{Y}}_j:=\left[\hat{\boldsymbol{y}}_{j,i}\right]_{v_i\in\mathcal{V}}\right\}_{\mathbf{f}_j\in\mathcal{F}}$, where $\hat{\boldsymbol{Y}}_j\in[0,1]^{20\times K}$ is the predicted label array on the entire dataset $\mathcal{V}$ by the model $\mathbf{f}_j$. Practice in ensemble learning has shown that a group of trained models would usually perform better than an individual model and could yield more reliable predictions (Zhou, 2012). Therefore, this study considers a soft voting mechanism to conclude the final node labels $\hat{\boldsymbol{Y}}:=\left[\hat{\boldsymbol{y}}_i\right]_{v_i\in\mathcal{V}},\hat{\boldsymbol{Y}}\in[0,1]^{20\times K}$, such that: $\hat{\boldsymbol{y}}_i=(\sum_{\mathbf{f}_j\in\mathcal{F}}p_j\hat{\boldsymbol{y}}_{j,i})/(\sum_{\mathbf{f}_j\in\mathcal{F}}p_j)$, or in the matrix form, $\hat{\boldsymbol{Y}}=(\sum_{\mathbf{f}_j\in\mathcal{F}}p_j\hat{\boldsymbol{Y}}_j)/(\sum_{\mathbf{f}_j\in\mathcal{F}}p_j)$, where $\hat{\boldsymbol{Y}}$ is a weighted average of the label arrays by all models whose column-sum pertains 2 for each post, and the weight p_j is the general performance score (e.g., accuracy, which will be discussed in Section 3.2.2) of model $\mathbf{f}_j$ on validation set.

Furthermore, the confidence of model prediction and the agreement/coherence among the different models also provide information for the reliability of the predictions (Zhou and Li, 2010). The former is trivial as the model confidence on all data points $\boldsymbol{\kappa}^{\text{con}}:=\left[\kappa_i^{\text{con}}\right]\in[0,1]^{K\times1}$ could be defined as the sum of top-n entries of the label vectors divided by two (since the sum of each label vector $\hat{\boldsymbol{y}}_i$ is two, as defined in Eq. (4)). The latter is also trivial when only two models are

concerned since the agreement of two vectors could be easily computed with any distance measure (e.g., cosine similarity, Euclidean distance, Jaccard Index, and/or cross-entropy). When $|\mathcal{F}| > 2$, this effectively becomes a problem of measuring the general linear dependence of a group of vectors composing the array $\hat{\boldsymbol{Y}}_i := \left[\hat{\boldsymbol{y}}_{j,i}\right]_{\mathbf{f}_j \in \mathcal{F}}, \hat{\boldsymbol{Y}}_i \in [0,1]^{20\times|\mathcal{F}|}$ for each node v_i. Inspired by GeoMatt22 (2020-12-10), this study computes the model agreement on all data points $\boldsymbol{\kappa}^{\text{agr}} =: \left[\kappa_i^{\text{agr}}\right] \in [0,1]^{K\times 1}$ from the first singular value $\sigma_{\boldsymbol{Z}_i,1}$ of the centred (subtracted by row-means) and normalized (divided by vector lengths) label matrix $\boldsymbol{Z}_i := \left[\boldsymbol{z}_{j,i}/\|\boldsymbol{z}_{j,i}\|\right]_{\mathbf{f}_j \in \mathcal{F}}, \boldsymbol{z}_{j,i} = \hat{\boldsymbol{y}}_{j,i} - \sum_j \hat{\boldsymbol{y}}_{j,i}/|\mathcal{F}|$ based on its Singular Value Decomposition (SVD) results, so that:

$$\kappa_i^{\text{agr}} = (\sigma^2_{\boldsymbol{Z}_i,1} - 1)/(|\mathcal{F}| - 1). \tag{9}$$

This is effective since the first several singular values measure how much variance of the matrix could be explained by its low-rank approximation, which is equivalent to eigenvalues in Principal Component Analysis (PCA) in statistics. The value of $\boldsymbol{\kappa}^{\text{agr}}$ ranges theoretically from the largest possible value (i.e., 1) when there are $|\mathcal{F}|$ completely parallel vectors in $\boldsymbol{Z}_i$, to the smallest possible value (i.e., 0) when all vectors are orthogonal (under the condition that $|\mathcal{F}| < 20$).

2.3.4. Spatial diffusion of node labels

In order to map the predicted node labels on the topological/geographical space, the label array $\hat{\boldsymbol{Y}}$ computed in Section 2.3.3 is further aggregated spatially, going one step further than the research conducted in Liu and De Sabbata (2021), where the labels of individual post nodes were directly drawn on maps. In Bai et al. (2022), the mapping relations of the posts to spatial nodes are also provided. For a city, an undirected weighted graph $G = (V, E, \boldsymbol{W})$ denotes its geographical representation obtained from Open Street Map (Boeing, 2017), where $V = \{v_k\}, k \in [0, |V|)$ is the node set of spatial intersections in a walkable network, $(v_k, v_{k'}) \in E \subseteq V \times V$ is a link marking if two spatial nodes are reachable to each other within 20 min by all means of transportation, and $\boldsymbol{W} := \left[W_{k,k'}\right] \in [0,1]^{|V|\times|V|}$ is a non-negative weighted adjacency matrix whose diagonal entries $W_{k,k}$ are all 1, recording the temporal closeness (i.e., the shorter time it takes to travel, the closer this weight gets to 1) between any pair of nodes v_k and $v_{k'}$, where $W_{k,k'} = 0$ when the nodes are not connected (not reachable within 20 min). Moreover, $\boldsymbol{B} := \left[B_{i,k}\right] \in \{0,1\}^{K\times|V|}$ records the one-hot mapping relation from posts nodes $\mathcal{V}$ to spatial nodes V, effectively a binary bi-adjacency matrix of a bipartite graph $\mathcal{B} = (\mathcal{V}, V, \mathscr{E}, \boldsymbol{B})$ connecting both node sets, where $(v_i, v_k) \in \mathscr{E} \subset \mathcal{V} \times V$ marks the link if a post is located nearby a spatial node. Note that the following relationship holds according to Bai et al. (2022): $\boldsymbol{A}^{\text{SPA}} = \left(\boldsymbol{B}\boldsymbol{W}\boldsymbol{B}^{\mathsf{T}} > 0\right) = \boldsymbol{B}\left(\boldsymbol{W} > 0\right)\boldsymbol{B}^{\mathsf{T}} \in \{0,1\}^{K\times K}$.

Without loss of generality, the processes of spatially aggregating and diffusing the node labels are visualized in Fig. 3, taking the neighbours of a generic spatial node v_k in both the spatial graph G as $\mathcal{N}_G(v_k) := \{v_{k'} | (v_k, v_{k'}) \in E \text{ or } W_{k,k'} > 0\} \subset V$ and in the bipartite graph $\mathcal{B}$ as $\mathcal{N}_{\mathcal{B}}(v_k) := \{v_i | (v_i, v_k) \in \mathscr{E} \text{ or } B_{i,k} = 1\} \subset \mathcal{V}$. The procedure takes place in two consecutive steps:

- Aggregating the predicted soft labels of all the posts nearby a spatial node $\hat{\boldsymbol{Y}}_{\mathcal{N}_{\mathcal{B}}(v_k)} := \left[\hat{\boldsymbol{y}}_i\right]_{v_i \in \mathcal{N}_{\mathcal{B}}(v_k)}$ to get the spatial node label $\hat{\boldsymbol{y}}_k \in [0,1]^{20\times 1}$, forming a 2D array $\hat{\boldsymbol{\mathcal{Y}}} := \left[\hat{\boldsymbol{y}}_k\right], \hat{\boldsymbol{\mathcal{Y}}} \in [0,1]^{20\times|V|}$;
- Diffusing the labels of all the spatial nodes to their spatial neighbours $\hat{\boldsymbol{\mathcal{Y}}}_{\mathcal{N}_G(v_k)} := \left[\hat{\boldsymbol{y}}_{k'}\right]_{v_{k'} \in \mathcal{N}_G(v_k)}$ based on their proximity iteratively, and *vice versa*, to get the final label $\boldsymbol{y}_k \in [0,1]^{20\times 1}$, with the label array $\boldsymbol{\mathcal{Y}} := \left[\boldsymbol{y}_k\right], \boldsymbol{\mathcal{Y}} \in [0,1]^{20\times|V|}$.

For the first step, the aggregation process should consider not only the respective values of the neighbouring labels, but also their importance (how dominant is the value compared to all the other nodes), prediction confidence (how confident are models predicting the label vectors containing this value) and prediction agreement (how reliable is this value). As it highly resembles the graph pooling operations in GNN, inspirations have been taken from literature (Li et al., 2015; Lee et al., 2019; Knyazev et al., 2019; Ma and Tang, 2021) to use an attention-based computation on each label category channel (as one instance among the 11 OUV or 9 HA categories) $\hat{\boldsymbol{y}}_C := \hat{\boldsymbol{Y}}^{\mathsf{T}}\boldsymbol{e}_C, \hat{\boldsymbol{y}}_C \in [0,1]^{K\times 1}$ to filter and summarize the labels, where $\boldsymbol{e}_C \in \{0,1\}^{20\times 1}$ is a one-hot unit vector only marking its C_{th} entry as 1. The attention value $\boldsymbol{s}_C \in [0,1]^{K\times 1}$ of all nodes v_i for any label category channel C could be computed as:

$$\boldsymbol{s}_C = \frac{\exp\left(\hat{\boldsymbol{y}}_C \odot (\boldsymbol{\kappa}^{\text{con}})^{1/\phi} \odot (\boldsymbol{\kappa}^{\text{agr}})^{1/\gamma}\right)}{\mathbf{1}_{K\times 1}^{\mathsf{T}} \exp\left(\hat{\boldsymbol{y}}_C \odot (\boldsymbol{\kappa}^{\text{con}})^{1/\phi} \odot (\boldsymbol{\kappa}^{\text{agr}})^{1/\gamma}\right)}, \tag{10}$$

where $\boldsymbol{\kappa}^{\text{con}}$ and $\boldsymbol{\kappa}^{\text{agr}}$ are model-level confidence and agreement scores on each node computed in Section 2.3.3, $\phi, \gamma \in \mathbb{R}$ are respectively parameters to adjust the contribution of confidence and agreement in the attention computation, such that when they get larger, high values of $\boldsymbol{\kappa}$ will be pushed closer to 1, $\odot$ is an element-wise Hadamard multiplication of vectors and arrays, and $\mathbf{1}_{K\times 1}$ is a K-dimensional vector of all 1s. Note that $\boldsymbol{s}_C$ is a stochastic vector over all the nodes.

Concatenating vectors $\boldsymbol{s}_C^{\mathsf{T}}$ for all category channels vertically together, an attention-based weight matrix $\boldsymbol{S} \in [0,1]^{20\times K}$ is obtained. This is then used as the weight of label array $\hat{\boldsymbol{\mathcal{Y}}}$ during the aggregation operation:

$$\begin{aligned}
\hat{\boldsymbol{\mathcal{Y}}}' &:= \begin{bmatrix} \hat{\boldsymbol{\mathcal{Y}}}^{\text{OUV}}_{11\times|V|} \\ \hat{\boldsymbol{\mathcal{Y}}}^{\text{HA}}_{9\times|V|} \end{bmatrix} = \left(\left(\boldsymbol{S} \odot \hat{\boldsymbol{Y}}\right)\boldsymbol{B}\right) \oslash \left(\boldsymbol{S}\boldsymbol{B}\right), \\
\hat{\boldsymbol{\mathcal{Y}}} &= \begin{bmatrix} \hat{\boldsymbol{\mathcal{Y}}}^{\text{OUV}} \oslash \left(\mathbf{1}_{11\times 1}\mathbf{1}_{11\times 1}^{\mathsf{T}}\hat{\boldsymbol{\mathcal{Y}}}^{\text{OUV}}\right) \\ \hat{\boldsymbol{\mathcal{Y}}}^{\text{HA}} \oslash \left(\mathbf{1}_{9\times 1}\mathbf{1}_{9\times 1}^{\mathsf{T}}\hat{\boldsymbol{\mathcal{Y}}}^{\text{HA}}\right) \end{bmatrix}
\end{aligned} \tag{11}$$

where $\oslash$ is the element-wise Hadamard division of two arrays, and the outcome of any spatial node $\hat{\boldsymbol{y}}_k$ is effectively a special form of weighted-average of the label vectors of all its neighbours $\hat{\boldsymbol{Y}}_{\mathcal{N}_{\mathcal{B}}(v_k)}$, scaled differently by the attention matrix $\boldsymbol{S}$ on each label category channel C. Similar to $\hat{\boldsymbol{Y}}$, the array $\hat{\boldsymbol{\mathcal{Y}}}$ is also a stack of two column-stochastic arrays for the OUV and HA labels, respectively.

Once the initial spatial node labels $\hat{\boldsymbol{\mathcal{Y}}}$ are computed, they could be used as the input state of an iterative diffusion process at the second step, during which each spatial node obtains information from its spatial neighbours and updates its own label while being reminded of its original state, until the labels converge at a steady state. This process resembles the graph filtering operation in GNN (Hamilton et al., 2017; Ma and Tang, 2021; Wu et al., 2022). For each spatial node v_k, its initial label is $\hat{\boldsymbol{y}}_k^{(0)} = \hat{\boldsymbol{y}}_k$. Assume the label is $\hat{\boldsymbol{y}}_k^{(t)}$ at the t_{th} iteration, then its next state after a diffusion step could be described as:

$$\hat{\boldsymbol{y}}_k^{(t+1)} = (1-\alpha)\hat{\boldsymbol{y}}_k + \alpha\frac{\sum_{v_{k'}\in\mathcal{N}_G(v_k)} W_{k,k'}\hat{\boldsymbol{y}}_{k'}^{(t)}}{\sum_{v_{k'}\in\mathcal{N}_G(v_k)} W_{k,k'}}, \tag{12}$$

or in its matrix form:

$$\hat{\boldsymbol{\mathcal{Y}}}^{(t+1)} = (1-\alpha)\hat{\boldsymbol{\mathcal{Y}}} + \alpha\hat{\boldsymbol{\mathcal{Y}}}^{(t)}\left(\boldsymbol{W}\boldsymbol{D}^{-1}\right), \tag{13}$$

where $\boldsymbol{D}$ is a diagonal matrix each entry of which records the degree (row-sum or column-sum) of the weighted symmetrical matrix $\boldsymbol{W}$, $\boldsymbol{W}\boldsymbol{D}^{-1}$ is the column-normalized stochastic matrix of $\boldsymbol{W}$, $\hat{\boldsymbol{\mathcal{Y}}}^{(t)} := \left[\hat{\boldsymbol{y}}_k^{(t)}\right] \in [0,1]^{20\times|V|}$ is the label array at the t_{th} iteration, and $\alpha \in [0,1)$ is a parameter controlling the importance of neighbouring nodes in the diffusion process. Even though label array $\hat{\boldsymbol{\mathcal{Y}}}$ only needs to be computed once needless of iterating, the rules described in Eqs. (12) and (13) enforce the spatial nodes to remember its original state at each iteration step, which could be effectively understood as that the spatial node v_k is pulling information both from its spatial neighbours $\mathcal{N}_G(v_k)$ (the second term in the Equations) and from its bipartite post neighbours $\mathcal{N}_{\mathcal{B}}(v_k)$ (the first term in the Equations) simultaneously on two respective graphs G and $\mathcal{B}$.

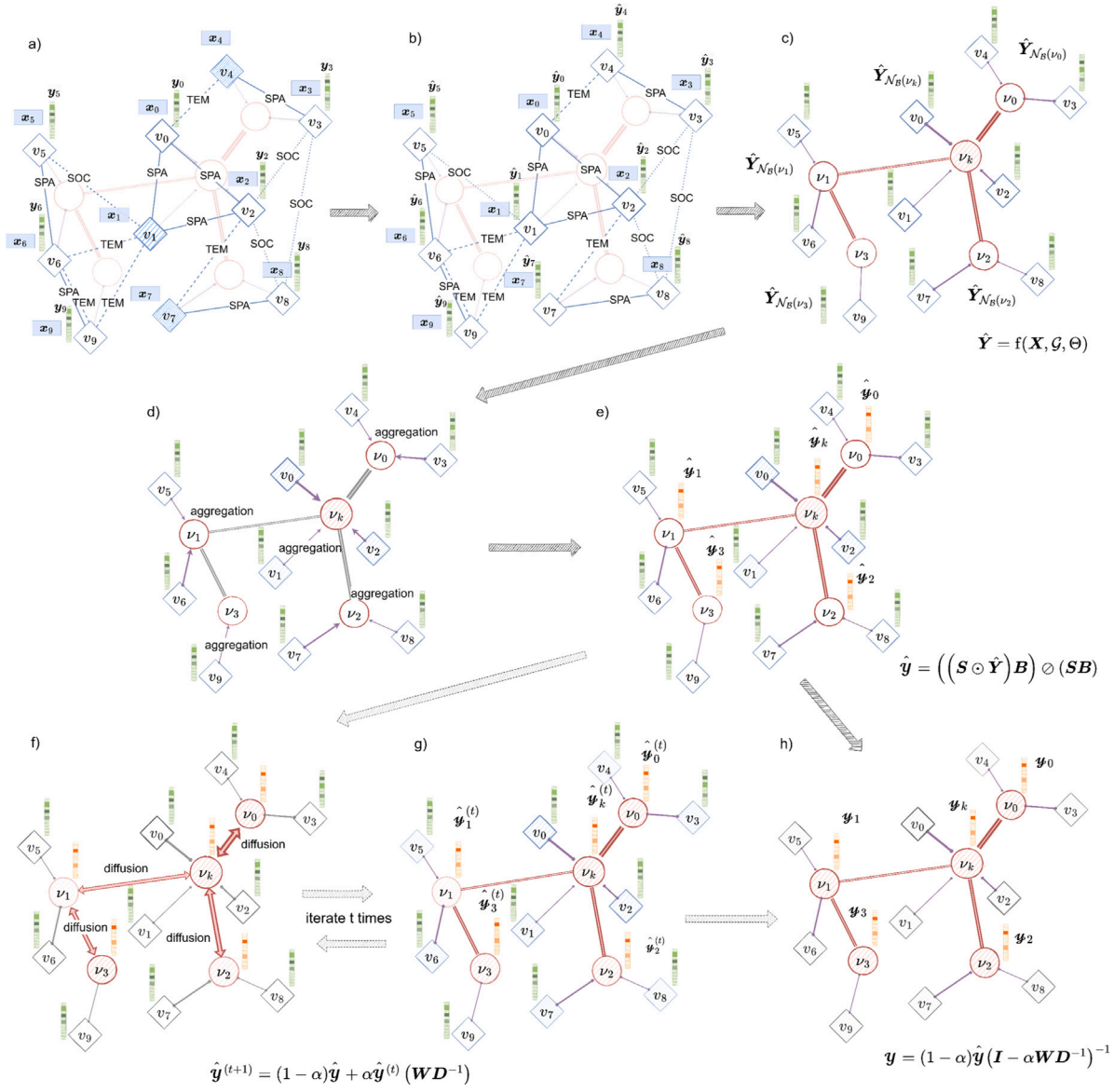

Fig. 3. The conceptually visualized semi-supervised learning, aggregation, and diffusion processes of node labels on a Post-level Attributed Multi-Graph (blue), a Post-Spatial Bipartite Graph (purple), and a Spatial Graph (red). Post nodes are represented with cylinders and spatial nodes with circles. (a) All posts are connected with temporal, spatial, or social links in a partially labelled attributed multi-graph, where each node has a complete feature array $\boldsymbol{x}_i$ and only some nodes have initial labels $\boldsymbol{y}_i$; (b) An estimated label vector $\hat{\boldsymbol{y}}_i$ is obtained for each post node with semi-supervised learning; (c) All posts neighbouring the spatial nodes v_k are labelled with $\hat{\boldsymbol{Y}}_{\mathcal{N}_B(v_k)}$; (d) Each spatial node aggregates (a single-sided process) the labels of neighbouring post nodes in the bipartite graph; (e) The initial label for each spatial node $\hat{\boldsymbol{y}}_k^{(0)} = \hat{\boldsymbol{y}}_k$ is obtained; (f) Each spatial node diffuses (a double-sided process) the labels of neighbouring spatial nodes in the spatial graph; (g) An intermediate state at step t of label diffusion on the spatial graph to obtain the label vector $\hat{\boldsymbol{y}}_k^{(t)}$; (h) The steady state when the spatial node label vector $\boldsymbol{y}_k$ converges. Note the iterative processes of (f) and (g) can be skipped by direct algebraic calculation in (h). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

For the steady state, the following equations hold:

$$\boldsymbol{\mathcal{Y}} = (1-\alpha)\hat{\boldsymbol{\mathcal{Y}}} + \alpha\boldsymbol{\mathcal{Y}}\left(\boldsymbol{W}\boldsymbol{D}^{-1}\right), \tag{14}$$

$$\boldsymbol{\mathcal{Y}}\left(\boldsymbol{I} - \alpha\boldsymbol{W}\boldsymbol{D}^{-1}\right) = (1-\alpha)\hat{\boldsymbol{\mathcal{Y}}}, \tag{15}$$

$$\text{therefore, } \boldsymbol{\mathcal{Y}} = (1-\alpha)\hat{\boldsymbol{\mathcal{Y}}}\left(\boldsymbol{I} - \alpha\boldsymbol{W}\boldsymbol{D}^{-1}\right)^{-1}. \tag{16}$$

For each row $\boldsymbol{y}_C^{\mathsf{T}} \in [0,1]^{1\times|V|}$ of $\boldsymbol{\mathcal{Y}}$ marking the distribution of one label category channel, the following also holds:

$$\boldsymbol{y}_C^{\mathsf{T}} = (1-\alpha)\hat{\boldsymbol{y}}_C^{\mathsf{T}}\left(\boldsymbol{I} - \alpha\boldsymbol{W}\boldsymbol{D}^{-1}\right)^{-1}, \tag{17}$$

where $\hat{\boldsymbol{y}}_C^{\mathsf{T}} := \boldsymbol{e}_C^{\mathsf{T}}\hat{\boldsymbol{\mathcal{Y}}}, \hat{\boldsymbol{y}}_C^{\mathsf{T}} \in [0,1]^{1\times|V|}$ is the C_{th} row of initial label array $\hat{\boldsymbol{\mathcal{Y}}}$. Note that the final array $\boldsymbol{\mathcal{Y}}$ is no longer a stack of two column-stochastic arrays respectively for OUV and HA labels since the sum of the "labels" of each spatial node can fluctuate around two, depending on the significance of the spatial nodes for each category channel. Also note that in the following equation:

$$\boldsymbol{y}_C = \left(\hat{\boldsymbol{y}}_C^{\mathsf{T}}(1-\alpha)\left(\boldsymbol{I} - \alpha\boldsymbol{W}\boldsymbol{D}^{-1}\right)^{-1}\right)^{\mathsf{T}} = \left((1-\alpha)\left(\boldsymbol{I} - \alpha\boldsymbol{W}\boldsymbol{D}^{-1}\right)^{-1}\right)^{\mathsf{T}}\hat{\boldsymbol{y}}_C, \tag{18}$$

the first component is clearly related to the generalized Katz Centrality (Benzi and Klymko, 2014; Zhan et al., 2017):

$$C_{\text{Katz}} = \beta\left(\boldsymbol{I} - \alpha\boldsymbol{A}^{\mathsf{T}}\right)^{-1}\mathbf{1}, \tag{19}$$

where the bias constant β is replaced with a constrained $1-\alpha$. Eq. (19) performs one more step of summation of Eq. (18) to obtain a centrality value. In other words, the calculation here uses an intermediate component of Katz centrality computation to weight the spatial labels (Nourian, 2016; Nourian et al., 2016; Zhan et al., 2017).

When $\alpha = 0$, no diffusion happens and the label vectors remain the same in all the steps. For Eqs. (16) and (17) to be solvable, the parameter α has to be chosen so that it is smaller than the reciprocal of the absolute value of the largest eigenvalue of $\boldsymbol{W}\boldsymbol{D}^{-1}$, i.e. $1/|\lambda|$, similar to the attenuation value for Katz Centrality computation. If this largest value is chosen, Eq. (19) becomes a standard eigenvector centrality (Gould, 1967; Bonacich, 1972). Moreover, by adjusting the local diffusion rule in Eqs. (12) and (13), the computation could be easily adjusted to other variants of spectral-based centrality such as PageRank (Page et al., 1999) and standard Katz Centrality (Katz, 1953). Note that the term $\hat{\boldsymbol{y}}_k^{(t)}$ denoting the last state of the nodes are not included in Eqs. (12) and (13). Appendix C will prove that adding such a term would end up calculating the same result in Eqs. (16) and (17) under certain constraints.

3. Experiments

3.1. Selected models and baselines

As described in Section 2.3.2, a group of models in a candidate set $\mathcal{F}$ will be trained on the datasets, and the best-performing model $\mathbf{f}_j$ of each type will be selected to output the model-specific predictions $\hat{\boldsymbol{Y}}_j$ to be further aggregated. To make the model ensemble various enough for its best effect (Zhou, 2012), the following diverse model types that are shown to be effective in literature are illustratively used:

Random classifier using prior distributions

- RDC - A Random Dummy Classifier baseline disregarding input features that generates random outputs based on the category distribution (prior) in the training set as shown in Table 1 (Baumer et al., 2015).

Graph-free classifiers using multi-modal features

- MLP - Multi-Layer Perceptron Classifiers with visual and textual features (Gardner and Dorling, 1998).

Homogeneous-graph GNN classifiers

- GCN - The Graph Convolution Network with initial residual connections and identity mapping (GCNII) proposed by Chen et al. (2020) as an extension for the vanilla GCN proposed by Kipf and Welling (2016).
- GAT - The Graph Attention Network proposed by Velickovic et al. (2017) with attention mechanism.
- GSA - Graph Sample and Aggregate (GraphSAGE) models proposed by Hamilton et al. (2017), which is especially effective for inductive learning, where knowledge learnt on one [sub-]graph is generalized across other unseen [sub-]graphs.

Heterogeneous-graph GNN classifiers

- HGSA - The heterogeneous GNN that handles each type of links separately with a different GraphSAGE sub-model, where results are aggregated when multiple types of links point to the same destination node (Zhang et al., 2019a).
- HGT - The Heterogeneous Graph Transformer proposed by Hu et al. (2020) that incorporates each type of links with an attention-based Transformer module (Vaswani et al., 2017).

During initial trials on the model structures, adding a linear layer in most graph-based models (except for GCN and GSA) and concatenating its output with that of the graph filters was found to boost the classification performance on *VEN* dataset. This is probably because the three types of links in *VEN*, i.e., the temporal, social, and spatial connections of the posts are all weak relations so that concatenating the neighbour features with the learnt feature of the node itself could overcome possible "over-smoothing" problem on these GNN, where individual features of all the nodes are forgotten and replaced by a universal aggregated one (Li et al., 2018). Also note that the Relational Graph Convolution Networks (Schlichtkrull et al., 2018) are not used as candidate models, as they assume that there only exists at most one type of relations between any two nodes, which is not the case in *VEN*, as two posts can be taken by the same person (socially similar) at the same place (spatially similar) in the same week (temporally similar).

3.2. Model training processes

3.2.1. Sub-sampling of graphs

The `NeighborLoader` in PyTorch Geometric (PyG) library (Fey and Lenssen, 2019), which is based on the Neighbour Sampler introduced by Hamilton et al. (2017), is used to generate sub-graphs $\mathcal{G}_s$ for all graph-based classifiers. A mini-batch of 32 post nodes are used as the input nodes $\mathcal{V}_{\text{batch}}$ for all sorts of subsets in $\mathcal{V}_{\text{train}}, \mathcal{V}_{\text{val}}, \mathcal{V}_{\text{test}}$, and $\mathcal{V}_{\text{unlab}}$. To make the GNN models compatible, for Homogeneous-graph GNN Classifiers (GCN, GAT, GSA), 75 neighbours are sampled for each node for two iterations, and for heterogeneous-graph GNN Classifiers (HGSA, HGT), 25 neighbours are sampled for each node and link type for two iterations. This effectively reduces the size of sub-graphs: the total number of links from the order of 1×10^6 in *VEN* and 1×10^8 in *VEN-XL* all to the order of 1×10^5 in the sub-graphs. This is especially desirable for datasets at scales such as *VEN-XL* for it to fit in computer memory during training and inference.

3.2.2. Evaluation metrics

Cross-Entropy of the soft labels are used as the loss functions $\ell_{\text{V}}, \ell_{\text{A}}$ for both OUV and HA classifications, while the parameter $\omega_{\text{V/A}}$ mentioned in Eq. (4) is set to 1 for simplicity during training.

For OUV classification, Top-1 Accuracy ($p^{\text{OUV(1)}}$), Top-n Accuracy ($p^{\text{OUV(n)}}$), and Order-n Jaccard Index ($p^{\text{OUV(nJ)}}$) are used as general evaluation metrics, while for HA classification, only Top-1 Accuracy ($p^{\text{HA(1)}}$) is used, since HA categories were assumed to be more precise in Bai et al. (2022). Let $\text{topk}(\boldsymbol{v}, n)$ denote a function returning an ordered set containing the indices of the top-n entries of a generic vector $\boldsymbol{v}$, then the evaluation metrics on any subset $\mathcal{V}_* \in \{\mathcal{V}_{\text{val}}, \mathcal{V}_{\text{test}}\}$ by model $\mathbf{f}_j$ can be respectively described as:

$$p_{*,j}^{\text{OUV(1)}} = \frac{\sum_{v_i \in \mathcal{V}_* \cap \mathcal{V}_{\text{V+,A-}}} \left(\text{topk}(\hat{\boldsymbol{y}}_{j,i}^{\text{OUV}}, 1) = \text{topk}(\boldsymbol{y}_i^{\text{OUV}}, 1)\right)}{\left|\mathcal{V}_* \cap \mathcal{V}_{\text{V+,A-}}\right|} \tag{20}$$

$$p_{*,j}^{\text{OUV(n)}} = \frac{\sum_{v_i \in \mathcal{V}_* \cap \mathcal{V}_{\text{V+,A-}}} \left(\text{topk}(\hat{\boldsymbol{y}}_{j,i}^{\text{OUV}}, 1) \in \text{topk}(\boldsymbol{y}_i^{\text{OUV}}, n)\right)}{\left|\mathcal{V}_* \cap \mathcal{V}_{\text{V+,A-}}\right|} \tag{21}$$

$$p_{*,j}^{\text{OUV(nJ)}} = \frac{\sum_{v_i \in \mathcal{V}_* \cap \mathcal{V}_{\text{V+,A-}}} \frac{\left|\left(\hat{y}_{j,i}^{\text{OUV}} > \frac{1}{n+1}\right) \wedge \left(y_i^{\text{OUV}} > \frac{1}{n+1}\right)\right|}{\left|\left(\hat{y}_{j,i}^{\text{OUV}} > \frac{1}{n+1}\right) \vee \left(y_i^{\text{OUV}} > \frac{1}{n+1}\right)\right|}}{\left|\mathcal{V}_* \cap \mathcal{V}_{\text{V+,A-}}\right|} \tag{22}$$

$$p_{*,j}^{\text{HA(1)}} = \frac{\sum_{v_i \in \mathcal{V}_* \cap \mathcal{V}_{\text{V-,A+}}} \left(\text{topk}(\hat{\boldsymbol{y}}_{j,i}^{\text{HA}}, 1) = \text{topk}(\boldsymbol{y}_i^{\text{HA}}, 1)\right)}{\left|\mathcal{V}_* \cap \mathcal{V}_{\text{V-,A+}}\right|}, \tag{23}$$

where Eq. (22) computes the Intersection over Union (Jaccard Index) of two sets of indices pointing to vector entries with values larger than a threshold (e.g., when $n = 3$, the computation is about logits larger than .25), being an effective way of evaluating soft label classification.

Furthermore, the per-class metrics of precision, recall, F1 score (harmonic average of precision and recall), and confusion matrix are used to inspect the model performance on each OUV and HA category channel. Moreover, since *VEN* and *VEN-XL* are unbalanced datasets as mentioned in Section 2.2 where some small classes only exist in top-n rather than top-1 labels, they are never counted in per-class metrics calculation as "true-positive" instances. As an explorative treatment, top-n per-class metrics are computed with the Algorithm 1, where the predicted and "ground-truth" top-n classes are permuted to obtain n^2 confusion matrices, which are further summed and normalized. Note

the diagonal entries of normalized confusion matrix $\tilde{\boldsymbol{M}}$ are effectively top-n F1 scores of top-n precision and recall. A similar explanation applies to the off-diagonal entries.

Algorithm 1: Computing Top-n Per-Class Metrics

Data: Number of Classes N, $1 \leq n \leq N$, a $N \times K$ Label Array $\boldsymbol{Y}$, a $N \times K$ Predicted Label Array $\hat{\boldsymbol{Y}}$, Standard Confusion Matrix Function of Index Arrays $\mathrm{ConfMat}(\boldsymbol{d}, \hat{\boldsymbol{d}})$

Result: Normalised Top-n Confusion Matrix $\tilde{\boldsymbol{M}}$, Top-n Precision $\boldsymbol{p}$, Top-n Recall $\boldsymbol{r}$, Top-n F1 Score $\boldsymbol{f}$

$\epsilon \leftarrow 0.0000001$;
$i, j, l, m \leftarrow 0$;
$\boldsymbol{M}, \tilde{\boldsymbol{M}} \leftarrow N \times N$ arrays of 0s;
$\boldsymbol{D}, \hat{\boldsymbol{D}} \leftarrow K \times n$ arrays of 0s;
$\mathbf{v}, \boldsymbol{p}, \boldsymbol{r}, \boldsymbol{f} \leftarrow N \times 1$ arrays of 0s;
$\mathbf{d}, \hat{\mathbf{d}} \leftarrow K \times 1$ arrays of 0s;
$\boldsymbol{D} \leftarrow \mathrm{topk}(\boldsymbol{Y}, n)$;
$\hat{\boldsymbol{D}} \leftarrow \mathrm{topk}(\hat{\boldsymbol{Y}}, n)$; `/Indices of top-n entries`
for $i \in [0, n)$ **do**
  $\boldsymbol{d} \leftarrow \boldsymbol{D}[:, i]$; `/Indices of` i_{th} `largest entries`
  for $j \in [0, n)$ **do**
    $\hat{\boldsymbol{d}} \leftarrow \hat{\boldsymbol{D}}[:, j]$;
    $\boldsymbol{M} \leftarrow \boldsymbol{M} + \mathrm{ConfMat}(\boldsymbol{d}, \hat{\boldsymbol{d}})$;
  end
end
$\boldsymbol{v} = \boldsymbol{M}.\mathrm{diagonal}()$; `/The diagonal entries`
for $l \in [0, N)$ **do**
  $\boldsymbol{p}[l] \leftarrow \boldsymbol{v}[l] / (\boldsymbol{M}[l, :].\mathrm{sum}() - (n-1) \times \boldsymbol{v}[l] + \epsilon)$;
  $\boldsymbol{r}[l] \leftarrow \boldsymbol{v}[l] / (\boldsymbol{M}[:, l].\mathrm{sum}() - (n-1) \times \boldsymbol{v}[l] + \epsilon)$;
  $\boldsymbol{f}[l] \leftarrow 2 \times \boldsymbol{p}[l] \times \boldsymbol{r}[l] / (\boldsymbol{p}[l] + \boldsymbol{r}[l] + \epsilon)$;
  for $m \in [0, N)$ **do**
    $\tilde{\boldsymbol{M}}[l, m] = 2 \times \boldsymbol{M}[l, m] / (\boldsymbol{M}[l, :].\mathrm{sum}() + \boldsymbol{M}[:, m].\mathrm{sum}() - 2 \times (n-1) \times \boldsymbol{M}[l, m] + \epsilon)$;
  end
end

3.2.3. Implementations of experiments

As briefly described in Section 2.3.2, the training procedure consists of the following steps: (1) for each model type, hyper-parameter searching was performed on sampled sub-graphs of *VEN* for 300–1000 epochs of training on $\mathcal{V}_{\mathrm{train}}$ with grid search in small ranges, where early-stopping was implemented based on the overall performance on validation set $\mathcal{V}_{\mathrm{val}}$; (2) the hyper-parameter configuration of the selected best models are used to re-train model checkpoints to be stored and used for inference; (3) the stored models are evaluated with metrics mentioned in Section 3.2.2 on both validation set $\mathcal{V}_{\mathrm{val}}$ and test set $\mathcal{V}_{\mathrm{test}}$ with 10 runs of different random seeds since some GPU-based models do not generate exactly same outcomes given a fix random seed; (4) once the overall performance of a model type is acceptable, it is used to predict the final label arrays $\hat{\boldsymbol{Y}}_j$ on the entire dataset $\mathcal{V}$ to be further aggregated; (5) Instead of repeating the same training process for *VEN-XL*, the model checkpoints obtained in step 2 are directly evaluated with $\mathcal{V}_{\mathrm{train}}$, $\mathcal{V}_{\mathrm{val}}$ and $\mathcal{V}_{\mathrm{test}}$ of *VEN-XL* (all practically test sets) and used to predict label arrays since it is assumed that the model checkpoints are generalizable in inductive learning. All models are implemented using building blocks provided by PyTorch Geometric (PyG) library. The datasets are structured and stored respectively as `Data` and `HeteroData` classes in PyG for different model types. More details of the training settings can be found in Appendix A.

To aggregate the predicted label arrays and perform SVD for the agreement score $\boldsymbol{\kappa}^{\mathrm{agr}}$, PyTorch (Paszke et al., 2019) is used. The sum of Top-1 HA Accuracy and Order-3 OUV Jaccard Index on both validation and test sets are used as the weight p_j for aggregation. To compute the confidence score $\boldsymbol{\kappa}^{\mathrm{con}}$, the top-4 entries of the aggregated label array $\hat{\boldsymbol{Y}}$ are used. For simplicity, parameters ϕ, γ in Eq. (10) are both set to 2 to compute the attention array $\boldsymbol{S}$. As for the spatial diffusion process, the parameter $\alpha \in [0, \min(1/|\lambda|, 1))$ is tested with 10 different values evenly dividing its theoretical lower and upper bounds (smaller than 1 for Eq. (13) to be meaningful) to test its effect on the distribution of the final label array $\hat{\boldsymbol{\mathcal{Y}}}$ on the spatial network.

3.3. Interpretation and visualization

3.3.1. Sensitivity on alternative conditions

To reflect on the assumption that graph-based models can better deal with semi-supervised learning tasks with a large proportion of missing features and/or labels, the trained model checkpoints are directly evaluated on an altered validation set $\mathcal{V}_{\mathrm{val}}$ where the visual or textual features of the mini-batches are masked and clipped to 0, while all the other nodes in the sampled graphs $\mathcal{G}_s$ are intact.

The usefulness of three link types $\{\boldsymbol{A}^{\mathrm{TEM}}, \boldsymbol{A}^{\mathrm{SPA}}, \boldsymbol{A}^{\mathrm{SOC}}\}$ are also experimented. For homogeneous graph models, the simple composed links $\boldsymbol{A}$ are replaced by each sub-link type to sample the sub-graphs for evaluation on $\mathcal{V}_{\mathrm{val}}$ in mini-batches. For heterogeneous graph models, only one link type is kept or masked during sub-graph sampling, yielding six different alternative performance scores on $\mathcal{V}_{\mathrm{val}}$.

As an alternative to the original graph links provided by Bai et al. (2022), a feature-based k-Nearest Neighbour (kNN) graph structure is also tested for homogeneous graph models. Since textual features have missing values, only visual features $\boldsymbol{X}^{\mathrm{vis}}$ are used to compute an adjacency matrix $\boldsymbol{A}^{\mathrm{kNN}} \in \{0, 1\}^{K \times K}$, where each entry $A^{\mathrm{kNN}}_{i,i'} = 1$ only if the post node $v_{i'}$ is within the 3 nearest neighbours of v_i based on cosine similarity. The kNN graph structure is computed with the `knn_graph` function of PyG library.

3.3.2. Interpreting the association of input features

For the final post-level label array $\hat{\boldsymbol{Y}}$ and the initial spatial-level label array $\hat{\boldsymbol{\mathcal{Y}}}$ before diffusion, rectangular co-occurrence matrices $\boldsymbol{O} \in \mathbb{N}^{11 \times 9}$ of top-3 OUV and top-1 HA categories are computed, where each matrix entry is normalized by dividing the total number of examples used for computation. When computing $\boldsymbol{O}$ for post-level label $\hat{\boldsymbol{Y}}$, only the posts whose sum of confidence score $\boldsymbol{\kappa}^{\mathrm{con}}$ and agreement score $\boldsymbol{\kappa}^{\mathrm{agr}}$ were above the 25% quantile are considered. These matrices can be used to explain the association of OUV and HA categories as well as their general distributions. When two categories from OUV and HA have high co-occurrence, they could be well-associated, informative for further heritage study investigations.

Furthermore, GNNExplainer (Ying et al., 2019) is illustratively used for GAT and GSA on $\mathcal{V}_{\mathrm{train}}, \mathcal{V}_{\mathrm{val}}, \mathcal{V}_{\mathrm{test}}$ to compute the relative importance of all visual and textual features for each OUV and HA category, among which 473 features out of 1753 are more explainable with physical meanings, e.g., scene categories (Zhou et al., 2017), SUN attribute categories (Patterson et al., 2014), number of faces (Schroff et al., 2015), and origin of languages. For all nodes considered, GNNExplainer predicted the relative importance of all features for classifying each node in sampled sub-graph mini-batches for 200 epochs. The explainable features mentioned above that entered the top-250 rankings by each node are counted for each OUV and HA category. A bipartite graph connecting the features with the categories is visualized in Gephi with Force Atlas algorithm (Bastian et al., 2009; Jacomy et al., 2014), which could be considered as an interpretable lexicon of the cultural significance categories.

3.3.3. Statistical tests and spatial mapping

T-Tests and Analysis of Variance (ANOVA) are conducted on the difference of model performance, confidence scores, and agreement scores between datasets *VEN* and *VEN-XL* and among subsets $\mathcal{V}_{\mathrm{train}}, \mathcal{V}_{\mathrm{val}}, \mathcal{V}_{\mathrm{test}}, \mathcal{V}_{\mathrm{unlab}}$ to check the coherence and consistency of trained models. All statistical tests are performed with Pingouin library (Vallat, 2018).

Table 3
The performance (%) of each model type in *VEN* dataset on validation and test sets as mean±standard deviation, computed using the stored model checkpoints with ten runs of evaluation with different random seeds. The best two models on each metric are marked in bold.

Model	$p_{val}^{OUV(1)}$	$p_{test}^{OUV(1)}$	$p_{val}^{OUV(3)}$	$p_{test}^{OUV(3)}$	$p_{val}^{OUV(3J)}$	$p_{test}^{OUV(3J)}$	$p_{val}^{HA(1)}$	$p_{test}^{HA(1)}$
RDC	18.79 ± 3.12	18.75 ± 3.08	57.14 ± 2.19	56.46 ± 3.69	21.92 ± 1.16	22.67 ± 1.85	17.56 ± 1.67	18.09 ± 1.15
MLP[a]	80.79 ± 0.00	80.21 ± 0.00	**99.51 ± 0.00**	99.48 ± 0.00	75.79 ± 0.00	74.13 ± 0.00	**98.98 ± 0.00**	**98.21 ± 0.00**
GCN-kNN[a]	74.38 ± 0.00	72.92 ± 0.00	**99.51 ± 0.00**	**98.44 ± 0.00**	69.21 ± 0.00	68.40 ± 0.00	91.87 ± 0.00	97.38 ± 0.00
GAT	80.39 ± 0.43	**82.55 ± 0.42**	**99.51 ± 0.00**	**99.48 ± 0.00**	76.32 ± 0.21	**76.11 ± 0.29**	**98.07 ± 0.10**	97.38 ± 0.08
GSA	80.69 ± 0.72	79.06 ± 0.65	**99.51 ± 0.15**	**99.48 ± 0.00**	**77.17 ± 0.38**	75.48 ± 0.49	95.71 ± 0.21	97.08 ± 0.22
HGSA	**84.73 ± 1.14**	77.86 ± 0.35	99.11 ± 0.20	99.11 ± 0.33	**77.33 ± 0.60**	71.74 ± 0.42	96.63 ± 0.24	95.65 ± 0.30
HGT[a]	79.31 ± 0.00	78.65 ± 0.00	98.03 ± 0.00	**99.48 ± 0.00**	73.81 ± 0.00	74.05 ± 0.00	96.95 ± 0.00	96.42 ± 0.00
Aggregated	**84.23**	**81.77**	99.01	**100.00**	76.77	**76.30**	97.56	**98.21**

[a]Deterministic outputs on GPU by the stored model checkpoint with different random seeds.

Table 4
The performance (%) of each model type in *VEN-XL* dataset on train, validation, and test sets, computed directly using the stored model checkpoints trained on *VEN* as inductive learning setting. The best two models on each metric are marked in bold. OUV selection criteria are shortened as "V", and Heritage Attributes as "A".

Model	$p_{train}^{V(1)}$	$p_{val}^{V(1)}$	$p_{test}^{V(1)}$	$p_{train}^{V(3)}$	$p_{val}^{V(3)}$	$p_{test}^{V(3)}$	$p_{train}^{V(3J)}$	$p_{val}^{V(3J)}$	$p_{test}^{V(3J)}$	$p_{train}^{A(1)}$	$p_{val}^{A(1)}$	$p_{test}^{A(1)}$
MLP	79.16	80.53	80.52	**98.67**	98.70	**98.86**	74.42	75.25	75.22	91.58	**96.86**	**96.79**
GCN-kNN	76.01	75.54	76.43	96.80	96.67	96.53	70.65	71.65	71.67	85.93	91.41	91.24
GAT	**80.04**	**80.88**	**80.90**	98.47	**98.72**	98.61	**74.09**	73.50	73.44	**93.32**	96.28	96.01
GSA	75.92	78.19	78.21	98.44	98.69	98.37	72.73	**75.55**	**75.28**	90.09	94.69	94.10
HGSA	77.12	78.81	78.48	98.49	98.41	98.41	70.63	70.53	69.85	90.66	95.10	94.62
HGT	77.58	78.34	78.92	97.95	98.04	98.20	72.66	72.48	72.39	91.36	95.40	95.25
Aggregated	**80.54**	**81.49**	**81.81**	**98.67**	**98.77**	**98.83**	**75.93**	**76.57**	**76.45**	**91.62**	**96.54**	**96.11**

For each category channel of the final spatial label array $\boldsymbol{y}_C$ with each value of α as in Eq. (17), the global Moran's I is computed as the spatial auto-correlation measure (Moran, 1950; Rogerson and Sun, 2001; Rogerson, 2021) of each OUV and HA category, showing the effect of spatial diffusion on the final label distribution, such that:

$$I_C = \frac{|V|(\boldsymbol{y}_C - \bar{y}_C\mathbf{1})^\top \boldsymbol{W}(\boldsymbol{y}_C - \bar{y}_C\mathbf{1})}{\mathbf{1}^\top\boldsymbol{W}\mathbf{1} \times (\boldsymbol{y}_C - \bar{y}_C\mathbf{1})^\top(\boldsymbol{y}_C - \bar{y}_C\mathbf{1})}, \tag{24}$$

where $\mathbf{1}$ is a $|V|$-dimensional vector of all 1s, $\bar{y}_C$ is the mean of vector $\boldsymbol{y}_C$, and $\boldsymbol{W}$ is the spatial closeness matrix mentioned in Section 2.3.4, thus not a conventional weight matrix with zero diagonal entries (Chen, 2021).

The spatial clustering effect of hot spots (clusters of high values) of each category channel is found with the computation of local Moran's I and the simulated p values based on random re-assignment of values on the spatial nodes (Anselin, 1995; Rogerson and Sun, 2001), such that:

$$\boldsymbol{I}_C = (\boldsymbol{y}_C - \bar{y}_C\mathbf{1}) \odot \boldsymbol{W}(\boldsymbol{y}_C - \bar{y}_C\mathbf{1}). \tag{25}$$

The spatial statistics global and local Moran's I are computed using the ESDA: Exploratory Spatial Data Analysis tool of PySAL library (Rey and Anselin, 2007) with doubly-standardized weight transformation together with 9999 permutations to generate simulated distributions for estimating two-tailed p values with Bonferroni correction (Vander-Weele and Mathur, 2019), where all the other parameters are kept as default. This computation would return the same results as implementing Eqs. (24) and (25). Afterward, the values of OUV and HA categories on spatial nodes are mapped using QGIS (QGIS Development Team, 2009).

4. Results

4.1. Classification performance

The classification performance of all the models is shown in Table 3 for *VEN* and in Table 4 for *VEN-XL*, while detailed performance curves of each model checkpoint during training can be found in Fig. A.1. The selected candidate models all performed reasonably well, as they all appeared in the best two instances at least once among the evaluation metrics on *VEN*, far exceeding the random classifier RDC. Note that only GCN selected is based on kNN graph structure mentioned in Section 3.3.1, since it performed better as will be shown in Fig. 7. Since different random seeds would change the configuration of sampled sub-graphs and the group of neighbours a node can learn from, the classification performance can be affected. However, except for the top-1 OUV accuracy for HGSA model, other variances are generally small. Furthermore, as the goal of this study is not to select the best model architecture, but to have stable and reliable performance, no single model was selected as the "final" one to predict labels. Rather, the aggregated prediction of all models was used in further steps. In *VEN*, aggregated prediction performed well in all evaluation metrics, either being among the best two models or performing considerably to the best ones. Yet in *VEN-XL* where models were directly evaluated without further training or fine-tuning, the aggregated prediction performed best for all metrics in all subsets. It is remarkable that GAT performed arguably the best among the individual models both in *VEN* and *VEN-XL*, suggesting that it has decent generalizability. Note the general performance of selected models including the aggregated prediction on all evaluation metrics dropped significantly from *VEN* to *VEN-XL* on their respective validation and test sets according to one-sided paired T-Test, $t(55) = 4.517, p < .0001$, yet the effect size of this drop is minimum (Cohen's d=0.096), suggesting that the knowledge learned on the small *VEN* dataset during training has been successfully transferred and generalized to the large unseen *VEN-XL* dataset.

The per-class metrics of OUV and HA categories by the aggregated prediction array $\hat{\boldsymbol{Y}}$ on both datasets can be seen in Tables 5 and 6, respectively. For most cultural OUV selection criteria except for Criterion (v) about Land-use and almost all HA categories except for Artificial Products, the aggregated prediction performed reasonably well in both *VEN* used for training, and *VEN-XL* completely new to the models. The poor performance of OUV Criteria (v), (viii), (ix) and HA category Artificial Products is clearly related to their scarce presence in the training set of *VEN* shown in Table 1, where the models had to learn the key features of a category using less than 10 examples. Specifically, even though there are a few training data labelled as Criteria (v)(ix)(x), no data from validation and test sets are labelled with them, thus resulting blanks ('–') in Table 5. Future data augmentation is

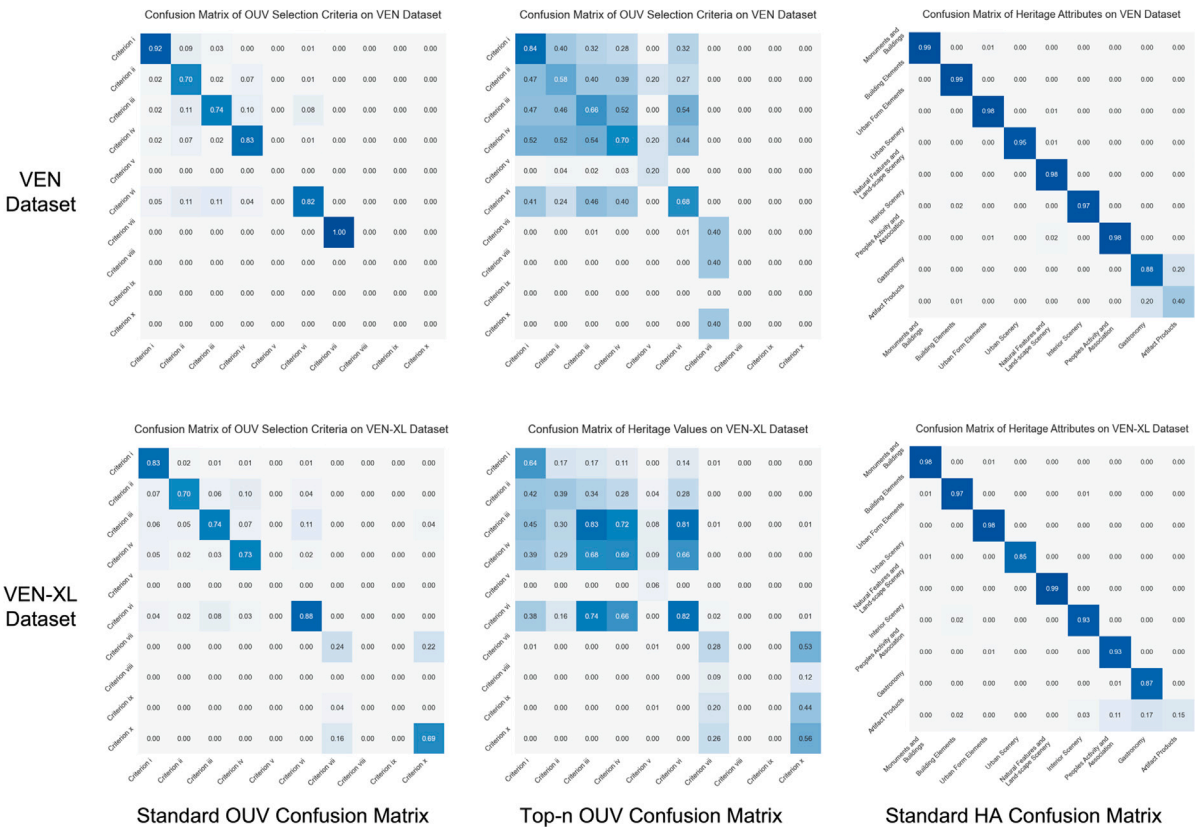

Fig. 4. The normalized top-1 and top-*n* confusion-matrix heatmaps of OUV selection criteria and Heritage Attributes classification of the aggregated prediction on both *VEN* and *VEN-XL* datasets. Note that these confusion matrices are not stochastic, and the entries represent the extent of confusion, where the diagonal entries are F1 scores in Tables 5 and 6.

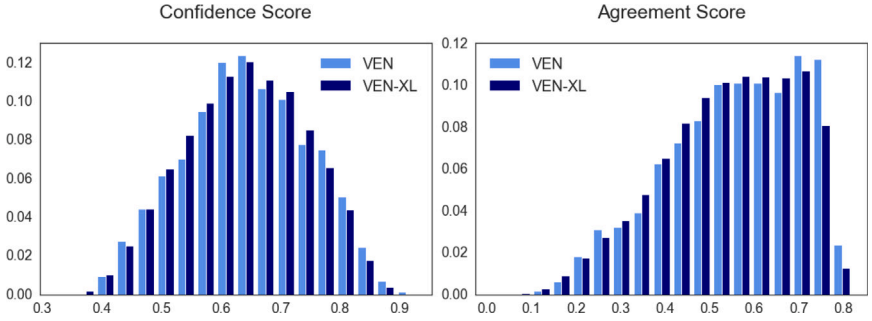

Fig. 5. The distribution of the confidence score κ^{con} and the agreement score κ^{agr} on both *VEN* (light blue) and *VEN-XL* (dark blue) datasets, both as density-based histograms. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

expected to teach the models specifically on these scarce classes. Under the same condition of scarcity, the prediction on Criterion (x) - Biodiversity, Urban Scenery, and Gastronomy performed remarkably well, suggesting that these classes are probably more clearly separated from the others in the feature space, easy for models to learn even with few-shot learning. The top-*n* per-class metrics proposed in Algorithm 1 is especially useful to evaluate scarce classes, as they may be absent as top-1 yet appear as top-*n* classes in validation and test sets, which can be seen in the cases of Criteria (v), (viii), (x) for *VEN* in Table 5. Such metrics are arguably stricter than standard per-class metrics in the sense that it evaluates the overlap of all top-*n* predictions with top-*n* labels (only when they are all the same, the metrics get to their theoretical maximum of 1), which could be seen as an extension of top-*n* accuracy with soft labels.

Moreover, the top-*n* per-class metrics allow a deeper observation of the confusion among the classes, as shown in Fig. 4. While Criterion (v) - Land Use is absent in standard OUV confusion matrices (for the same reason mentioned above that no data in validation and test sets of *VEN* has a top-1 label of it), the values in top-*n* confusion matrices give a hint on how other classes are confused and thus related with it: posts about land-use in Venice also concern with the influence of Venice to the world and its special architectural style near the canals. Posts concerning Criteria (iii), (iv), and (vi) are easily confused with each other, meaning that when people post about Venice on Flickr, themes about testimony of the past, architectural typology and the association of architectural and urban elements with human activity usually come together. The same goes for Criteria (vii) and (x) about natural beauty of the city and the living animals and plants indicating bio-diversity. For HA, Artifact Products can be confused with Gastronomy and People's Activity, which also makes sense as all three topics usually depict human and human-related objects. Such associations will be further elaborated in Section 4.3.

The confidence score κ^{con} and the agreement score κ^{agr} mentioned in Section 2.3.3 have similar distributions for *VEN* and *VEN-XL* datasets

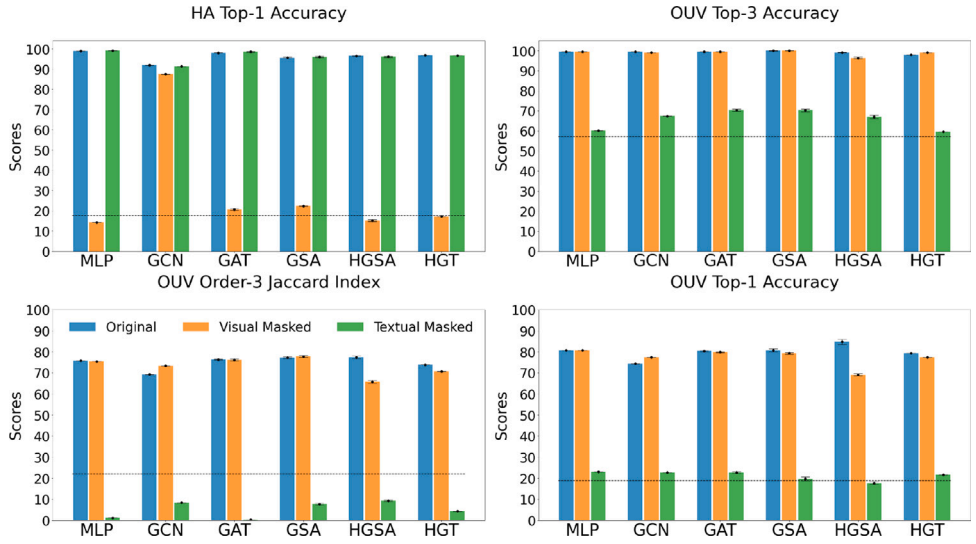

Fig. 6. The performance of all selected model checkpoints on the evaluation metrics when masking visual or textual features of mini-batches. The performance of the prior-based random classifier RDC in Table 3 is marked with dashed lines.

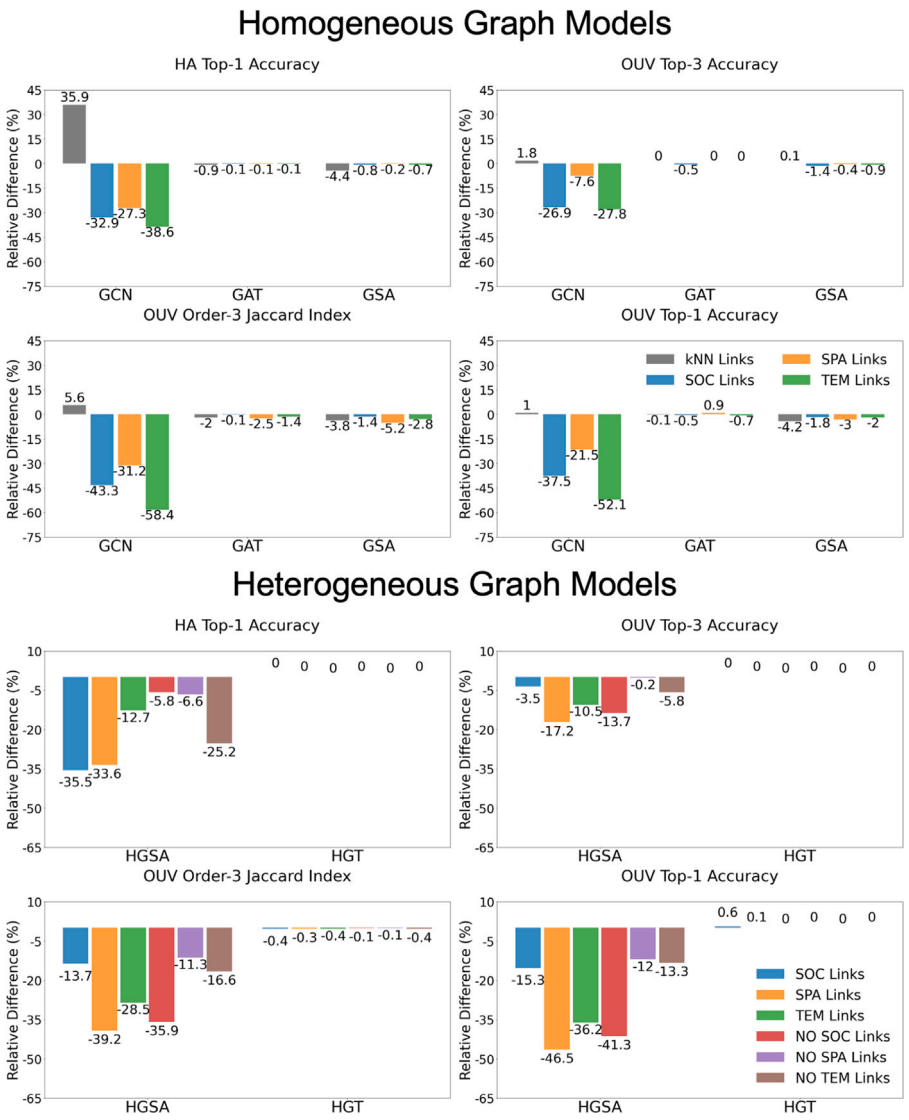

Fig. 7. The relative performance change of homogeneous and heterogeneous graph models directly evaluated on sub-graphs with one or two of the link types in $\{\boldsymbol{A}^{\text{TEM}}, \boldsymbol{A}^{\text{SPA}}, \boldsymbol{A}^{\text{SOC}}\}$, compared to the original composed links $\boldsymbol{A}$. The models with kNN links $\boldsymbol{A}^{\text{kNN}}$ were trained separately.

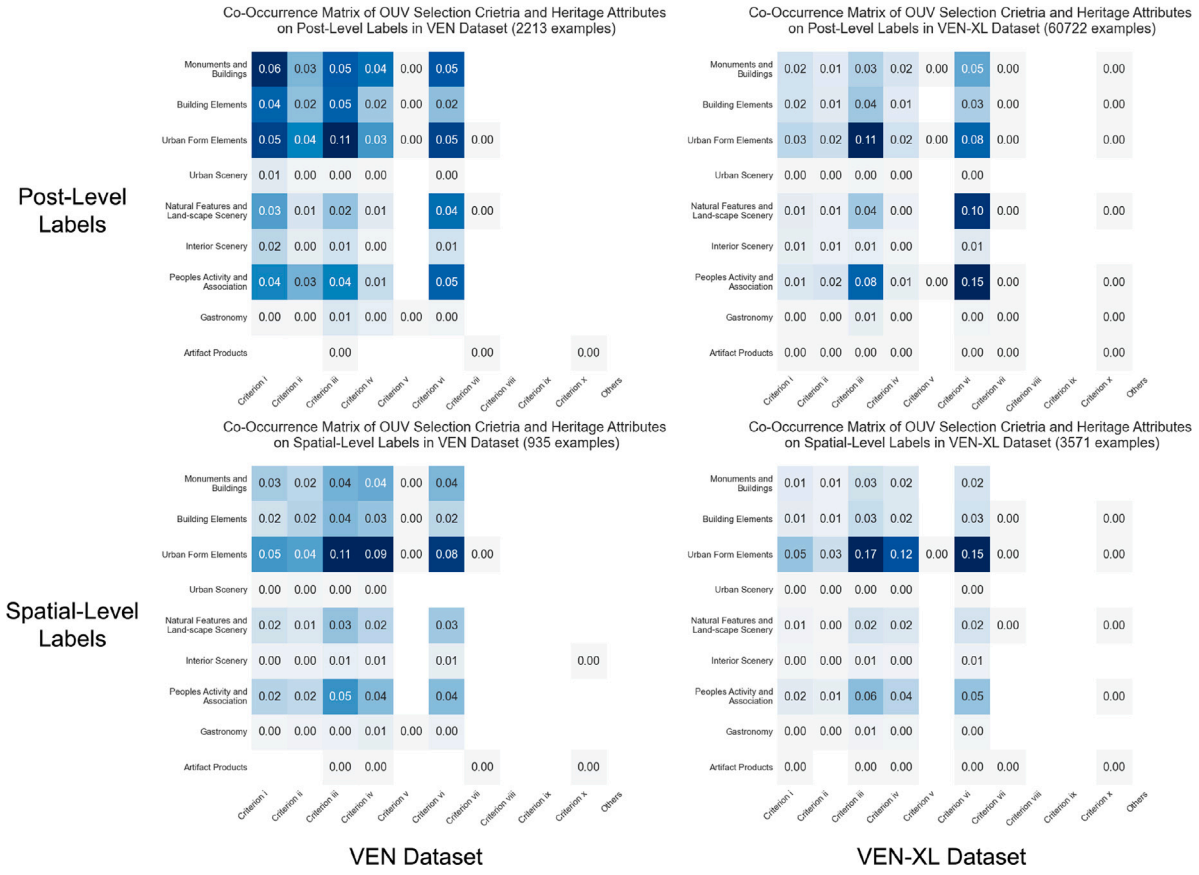

Fig. 8. The normalized co-occurrence matrix heatmaps $\boldsymbol{O}$ of the OUV and HA categories in post-level label array $\hat{\boldsymbol{Y}}$ and spatial-level label array $\hat{\boldsymbol{y}}$ in both *VEN* and *VEN-XL* datasets.

Table 5
The per-class performance metrics of OUV Selection Criteria classes in *VEN* and *VEN-XL* datasets. When no correct predictions were made for a class, the score would be 0.00; yet when no examples of a class were available, the score is marked as "–". The class "Others" is omitted since no examples were assigned to it.

Metrics	Precision	Recall	F1	Top-3 Precision	Top-3 Recall	Top-3 F1
Criterion (i) - Masterpiece	0.94 | 0.87	0.89 | 0.79	0.92 | 0.82	0.86 | 0.64	0.81 | 0.64	0.84 | 0.64
Criterion (ii) - Influence	0.76 | 0.59	0.65 | 0.87	0.70 | 0.70	0.63 | 0.27	0.53 | 0.74	0.58 | 0.39
Criterion (iii) - Testimony	0.68 | 0.69	0.80 | 0.79	0.74 | 0.74	0.73 | 0.93	0.61 | 0.74	0.66 | 0.83
Criterion (iv) - Typology	0.88 | 0.70	0.79 | 0.76	0.83 | 0.73	0.65 | 0.75	0.76 | 0.64	0.70 | 0.69
Criterion (v) - Land-use	– | 0.00	– | 0.00	– | 0.00	0.11 | 0.03	1.00 | 0.38	0.20 | 0.06
Criterion (vi) - Association	0.78 | 0.94	0.88 | 0.82	0.82 | 0.87	0.63 | 0.89	0.75 | 0.76	0.68 | 0.82
Criterion (vii) - Natural Beauty	1.00 | 0.16	1.00 | 0.55	1.00 | 0.24	0.25 | 0.17	1.00 | 0.94	0.40 | 0.28
Criterion (viii) - Geological Process	– | 0.00	– | 0.00	– | 0.00	0.00 | 0.00	0.00 | 0.00	0.00 | 0.00
Criterion (ix) - Ecological Process	– | 0.00	– | 0.00	– | 0.00	– | 0.00	– | 0.00	– | 0.00
Criterion (x) - Bio-diversity	– | 0.66	– | 0.73	– | 0.69	0.00 | 0.39	0.00 | 1.00	0.00 | 0.56

Table 6
The per-class performance metrics of heritage attributes classes in *VEN* and *VEN-XL* datasets.

Metrics	Precision	Recall	F1
Monument and buildings	0.99 | 0.98	0.99 | 0.98	0.99 | 0.98
Building elements	1.00 | 0.98	0.98 | 0.96	0.99 | 0.97
Urban form elements	0.99 | 0.99	0.98 | 0.97	0.98 | 0.98
Urban Scenery	0.91 | 0.74	1.00 | 1.00	0.95 | 0.85
Natural features and landscape scenery	0.99 | 0.99	0.97 | 0.99	0.98 | 0.99
Interior scenery	0.95 | 0.90	1.00 | 0.96	0.97 | 0.93
People's activity and association	0.96 | 0.99	1.00 | 0.88	0.98 | 0.93
Gastronomy	0.95 | 0.92	0.82 | 0.83	0.88 | 0.87
Artifact products	0.29 | 0.08	0.67 | 0.93	0.40 | 0.15

as visualized in Fig. 5. Two-way ANOVA F-Tests on the level of different datasets and on the level of zoomed-in subsets $\mathcal{V}_{\text{train}}$, $\mathcal{V}_{\text{val}}$, $\mathcal{V}_{\text{test}}$, and $\mathcal{V}_{\text{unlab}}$ is showed in Table 7. All effects are statistically significant, yet only the main effect of subset has large effect sizes η^2, and the main effect of the dataset and the interaction effect are all minimum, which can also be seen with Cohen's d computed with independent T-Tests with Welch's correction. The very small effect sizes on the level of dataset indicate that the significant drops of both scores from *VEN* to *VEN-XL* are mainly caused by the large sample size in *VEN-XL*, suggesting that the models function consistently and coherently in both datasets. Post hoc comparisons using the Tukey HSD test indicated that the scores in $\mathcal{V}_{\text{train}}$ are always significantly higher than $\mathcal{V}_{\text{val}}$, and $\mathcal{V}_{\text{test}}$, and the scores in $\mathcal{V}_{\text{unlab}}$ are always significantly lower than all the others with large effect size, while there is no significant difference between $\mathcal{V}_{\text{val}}$ and $\mathcal{V}_{\text{test}}$, as can be seen in Table B.1 of Appendix B. This

Table 7
Means, Standard Deviations, and Two-Way ANOVA Statistics on the Confidence and Agreement scores. An additional Independent T-Test with Welch's correction is also performed on the level of two datasets.

Score	*VEN*		*VEN-XL*		ANOVA				
	M	*SD*	*M*	*SD*	Effect	*df*	*F*	*p*	η^2
Confidence score κ^{con}									
–Training set $\mathcal{V}_{train}$	0.795	0.042	0.744	0.076	Dataset	1	59.938	<.0001	.0004
–Validation set $\mathcal{V}_{val}$	0.666	0.076	0.663	0.080	Subset	3	17,336.251	<.0001	.3827
–Test set $\mathcal{V}_{test}$	0.667	0.077	0.664	0.080	Dataset ×Subset	3	32.388	<.0001	.0012
–Unlabelled set $\mathcal{V}_{unlab}$	0.573	0.084	0.563	0.083	Residual	83,906			
(Overall)	0.644	0.105	0.638	0.102	$t(3158.402) = 2.910, p = .004$, Cohen's d=0.056				
Agreement score κ^{agr}									
–Training set $\mathcal{V}_{train}$	0.741	0.033	0.664	0.099	Dataset	1	110.854	<.0001	.0008
–Validation set $\mathcal{V}_{val}$	0.604	0.110	0.589	0.115	Subset	3	16,195.095	<.0001	.3662
–Test set $\mathcal{V}_{test}$	0.604	0.111	0.589	0.116	Dataset ×Subset	3	27.723	<.0001	.0001
–Unlabelled set $\mathcal{V}_{unlab}$	0.444	0.129	0.427	0.129	Residual	83,906			
(Overall)	0.556	0.152	0.541	0.149	$t(3160.154) = 5.235, p < .0001$, Cohen's d=0.100				

again shows the consistency and coherence of the model performance. When further aggregating the labels into spatial nodes, those posts with high prediction confidence and agreement (thus are more reliable) contribute more to attention score computation. Note the scores on the training set gets closer to the validation and test sets in *VEN-XL* than in *VEN* with lower means and larger standard deviations. This is probably because the models are not trained on *VEN-XL*, and the training set, therefore, becomes another [easier] validation/test set, as pointed out in Section 3.2.3.

4.2. Robustness of models

Fig. 6 shows the performance of selected models while masking the visual or textual features of the sub-sampled validation mini-batches. Masking visual features significantly lowers the HA scores, and masking textual features significantly lowers the OUV scores. This is a natural and consistent behaviour considering how those labels were originally derived: in Bai et al. (2022), HA labels were generated using images only and OUV labels were generated using texts only. In this study, however, the models have access to both textual and visual features when making classifications on both HA and OUV categories. GCN-kNN was the most robust model against the masking of visual features since the kNN graph structure $\boldsymbol{A}^{kNN}$ was computed before masking, unconsciously leaking the association information of visually similar images (and possibly their HA labels) to the models being trained. All graph-based models performed better than the graph-free MLP at HA classification while masking visual features, whereas the homogeneous models remained better than random classifier RDC. For OUV classification, Order-3 Jaccard Index of all models became extremely vulnerable and got far worse than RDC after masking textual features, since the requirement of being larger than $1/(n+1)$ in Eq. (22) cannot be easily fulfilled when models get uncertain of their predictions. Top-3 OUV Accuracy shows that almost all graph-based models (except for HGT) performed better than MLP (which was also better than RDC) while masking textual features, implying that those models managed to learn the missing textual information of a post from its neighbours, which is only possible on graphs. However, such an effect is not obvious for Top-1 OUV Accuracy, where most models performed only slightly better than RDC.

Fig. 7 shows the relative performance change of all graph-based models using different graph structures, compared to the original links. GCN trained on kNN graph $\boldsymbol{A}^{kNN}$ performed significantly better than the original links in all metrics, while GAT and GSA performed slightly worse on kNN graph, suggesting the necessity of using GCN-kNN as the selected candidate model in Tables 3 and 4. Changing graph structure only slightly lowers the performance on GAT and GSA, while not affecting HGT at all. This seems to suggest that these models work as long as there is some graph structure marking the relationship of data points, indifferent of the type of links. Meanwhile, GCN and HGSA are more dependent on the links used for inference.

The various behaviours imply that the selected models are divergent enough, suggesting that aggregating the prediction results to form an ensemble is both necessary and beneficial. The discussion on the complex effects of the model performance, however, falls out of the scope of this paper and invites further investigations in future studies.

4.3. Association of features and labels

Fig. 8 shows the co-occurrence matrices of OUV and HA categories as heatmaps, where frequent OUV-HA pairs imply the association of abstract OUV selection criteria and substantial Heritage Attributes. The four matrices on both post-level labels $\hat{\boldsymbol{Y}}$ and spatial-level labels $\hat{\boldsymbol{\mathcal{Y}}}$ in both *VEN* and *VEN-XL* datasets are similar to each other. The spatial-level distribution on *VEN-XL* is the most sparse (and concentrated) among the four matrices where most OUV-HA pairs focused on the large classes, i.e., Criteria (iii) and (vi) for OUV and Urban Form Element for HA. A similar yet more extreme pattern can be observed in Fig. B.1 in Appendix B when the parameter α gets larger, pushing the diffused spatial nodes label array $\boldsymbol{\mathcal{Y}}$ to a uniform-like distribution, suggesting possible "over-smoothing". A few OUV-HA pairs always stand out as associated categories in those co-occurrence matrices: (1) As the most common HA category, the Urban Form Elements always associate strongly with Criteria (iii), (iv) and (vi), suggesting that when people post about testimony of past, architecture type, and human-life-related traditions, they are usually immersed in the urban context of streets and squares; (2) The second largest HA category about People's Activity also associate strongly with Criteria (iii) and (vi), since they have obvious connections with human; (3) As expected, the most associated OUV category with Monuments and Building is Criterion (iv) about architecture typology, and that with Building Element is Criterion (iii) about testimony for a [possibly lost] tradition; (4) The most unexpected associations are the ones for Natural Features and Landscape Scenery, where the most relevant Criterion (vii) about natural beauty is always present but not in a dominant position, which has also been taken by Criterion (iii) and (vi). The pattern of OUV and HA category distribution will be further dis-aggregated and mapped spatially in Section 4.4 for detailed inspection.

Fig. 9 visualizes the explainable features that are shown to be important for classifying the nodes into each OUV and HA category, effectively forming a lexicon of features for the categories as a bi-partite graph. The contribution of features is interrelated to OUV/HA categories. For example, the recognized scene of "Canals in Urban

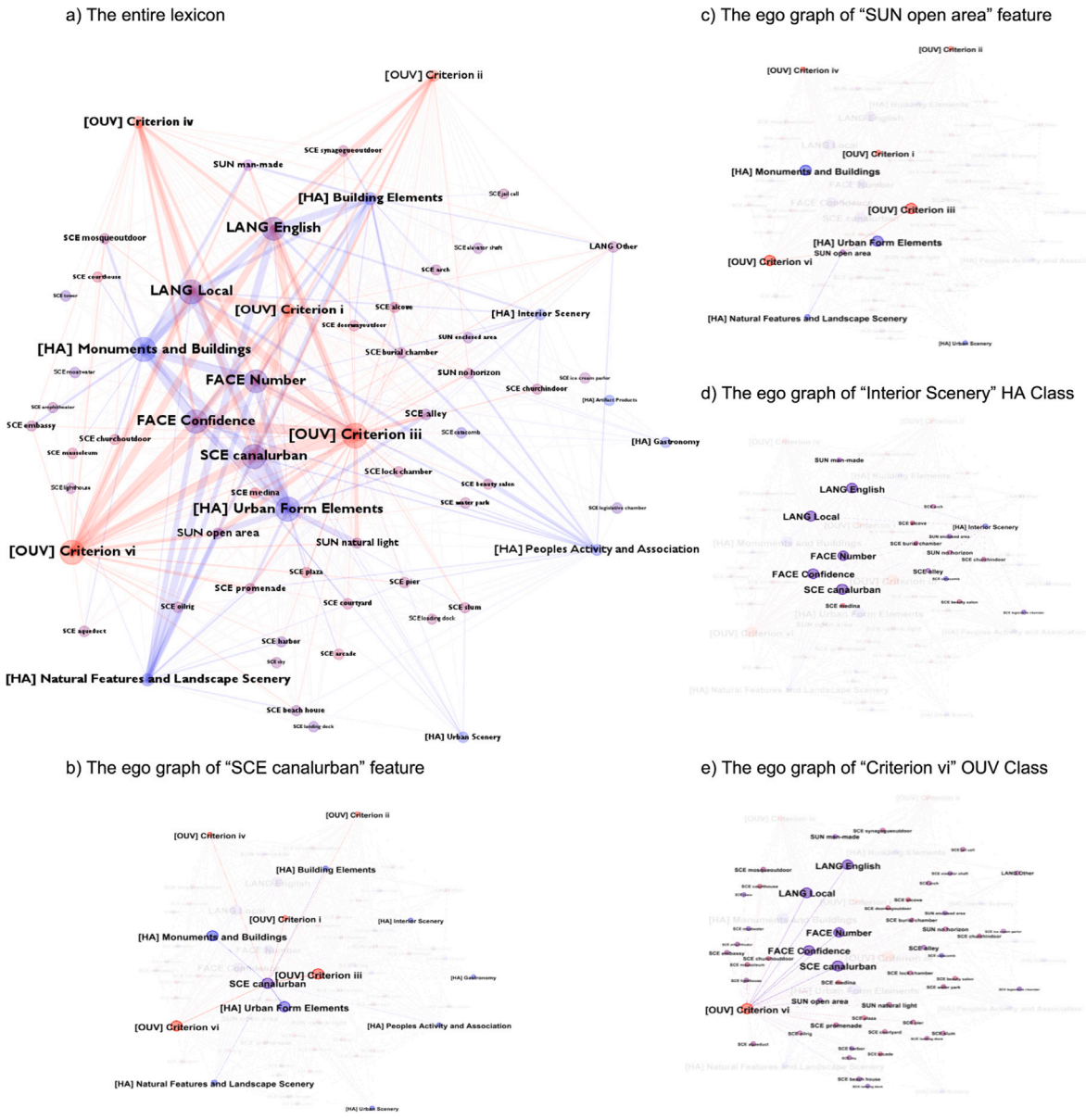

Fig. 9. The bipartite graph of feature nodes and OUV/HA category nodes showing the relative importance for explainable features while classifying the nodes belonging to each OUV and HA category. The larger a feature node is, the more this feature appeared in the top-250 important features while classifying a node based on GNNExplainer. The edge weights show the number of times the features contributed to the categories. Only nodes with a larger weighted degree of 8 are shown. Red lines are associations for OUV classes and blue lines for HA. Sub-figures b-e show ego graphs (a sub-graph of the entire lexicon in sub-figure a) around a specific feature or category node. “SCE” denotes scene category within Zhou et al. (2017); “SUN” denotes SUN attribute category in Patterson and Hays (2012); “LANG” denotes the detected language and “FACE” denotes face recognition results from Bai et al. (2022). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

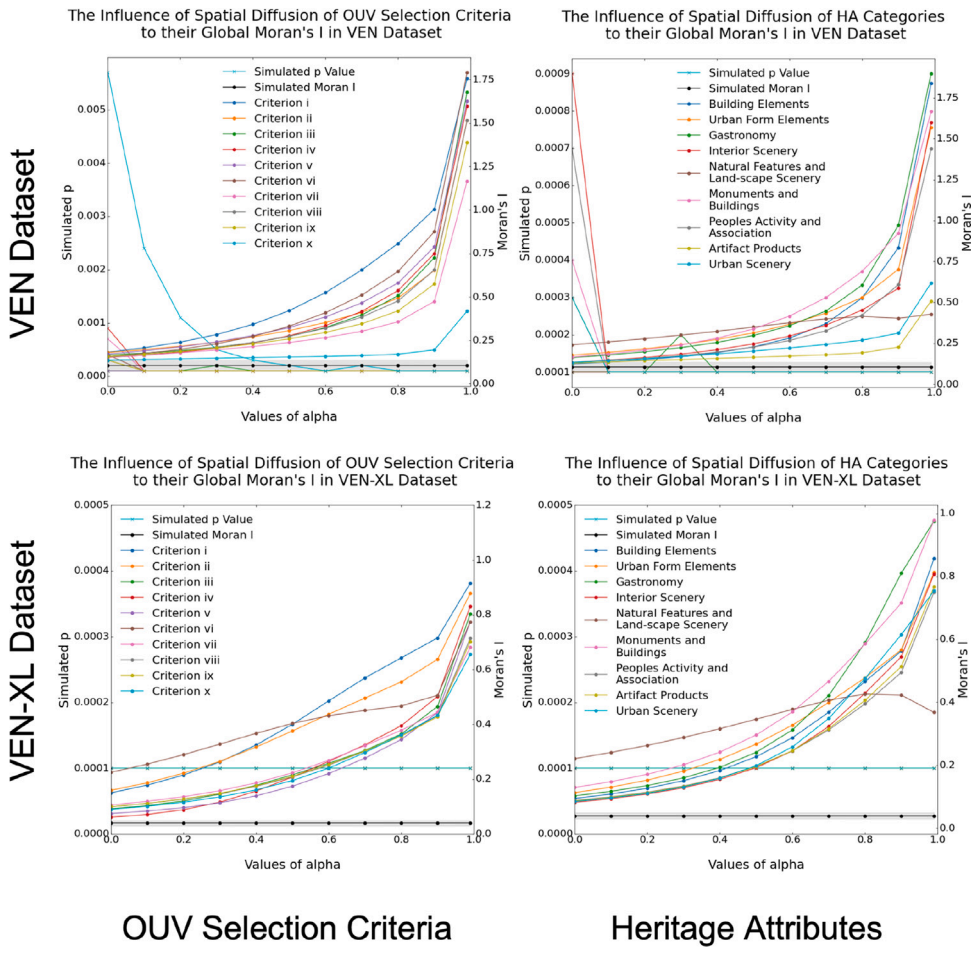

Fig. 10. The change of global Moran's I of each OUV and HA category when the diffusion parameter α changes in *VEN* and *VEN-XL*. A simulated distribution of expected values of I based on 9999 permutations is used to estimate the p values.

Environment" and the SUN attribute of "Open Area" from an image both contribute generally to almost all OUV/HA categories, especially on Criteria (iii)(vi) and "Urban Form Element", while "Open Area" has less to do with "Interior Scenery", "Building Elements", and "People's Activity and Association". While HA category "Interior Scenery" could be inferred with a limited range of features such as "Enclosed Area" and "Arch", OUV Criterion (vi) could be inferred from a large variety of visual and textual features, depending on the type of human activity taking place. The face recognition and language detection results appear to contribute universally to the classification of most categories, which could be possibly explained that the presence of human faces and the original languages of posts provide additional information that could not be inferred from features extracted with scene recognition models originally trained with images with few people and language models trained with English texts. However, among all visual and textual features, explainable ones are usually less informative than the higher-level hidden features, as can be seen in Fig. B.2. More concrete investigations are invited to explain this complex pattern.

4.4. Mapping of heritage cultural significance

Fig. 10 demonstrates that the global Moran's I for OUV and HA categories gradually increase as the diffusion parameter α ascends. For most categories in *VEN* and all in *VEN-XL*, a spatial auto-correlation is significant after Bonferroni correction ($p < .025/20$) even before diffusion compared to the permutated distributions, confirming the First Law of Geography. For smaller α values, the increases in Moran's I are not drastic, yet effectively further decrease the simulated p values. The largest value of $\alpha = 0.99$ yields extreme I values larger than 1 in *VEN*. This suggests that choosing a relatively small value for α could enhance the spatial pattern of the categories without disturbing their distributions too much. Note the expected value (mean) of I according to simulation is not the conventional $-1/(N-1)$, since the weight matrix $\boldsymbol{W}$ used here has non-zero diagonal entries and is not row-standardized. However, Fig. B.3 shows a similar pattern with the conventional weight matrix for computing Moran's I. The following sections will use $\alpha = 0.3$ for demonstrative purposes of exploratory spatial data analysis. The distribution of spatial node labels $\boldsymbol{\mathcal{Y}}$ in Fig. 11 also demonstrates a consistent pattern in *VEN* and *VEN-XL*: (1) five OUV and HA categories are relatively more dominant than the others; (2) the confidence of OUV labels for spatial nodes are generally lower than HA labels since OUV categories have to be sometimes inferred without textual information; (3) whereas the less dominant categories have lower means and quantile values, the "outliers" point to the exceptional spatial nodes representing specific OUV and HA categories. It further shows that although none of Criteria (vii) - (x) are inscribed with Venice in WHL, scarce cases related to Criteria (vii) and (x) can still be found.

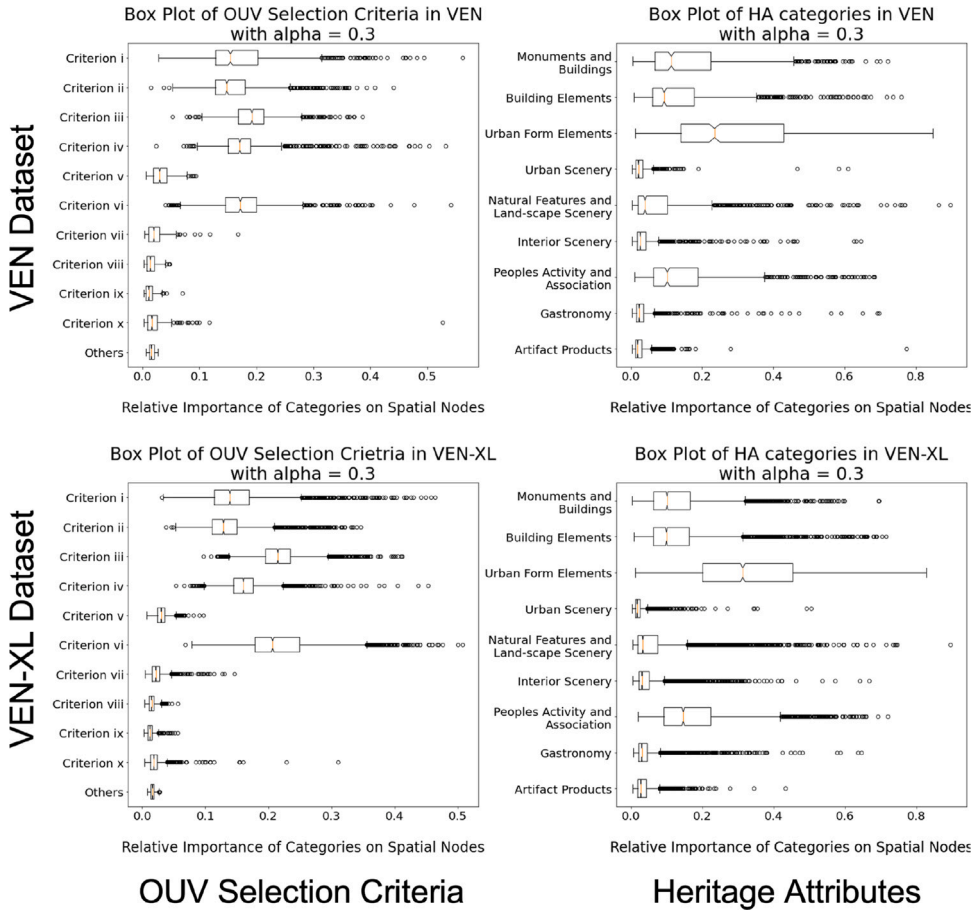

Fig. 11. The box plots of each OUV and HA category demonstrating the distributions of spatial node labels $\mathscr{Y}$ in both *VEN* and *VEN-XL* datasets.

Fig. 12 demonstrates the final maps of OUV and HA categories identified from Flickr showing their spatial distributions and auto-correlation patterns, together with illustrative examples. The magnitude of HA categories is generally higher than OUV, as also pointed out in Fig. 11. Almost all categories display spatial patterns of "hotspots" of high values appearing at nearby places, justified with significant local Moran's $\boldsymbol{I}$. Some categories are spread all over Venice, e.g., OUV Criterion (iii) about Testimony and HA Urban Form Elements, due to their universal nature, while others are much more concentrated at dedicated spots, e.g., OUV Criterion (iv) about Architecture Typology and HA People's Activity and Associations. Even though some categories are less present with far more limited range, e.g., OUV Criterion (v) about Land-Use and HA Artifact Product, the methodology does manage to find relevant spatial spots with posts of images and/or comments related to the topic. The OUV-HA pairs generally believed to associate with each other, such as Criterion (iv) about Architecture Typology and HA Monuments and Buildings, Criterion (vi) about Human Association and HA People's Activity and Associations, and Criterion (vii) about Natural Beauty and HA Natural Features and Landscape Scenery, partly overlap with each other, yet not totally identical, showing the nuances of the concepts reflected in social media posts. Interestingly, the hotspot visualization and illustrated examples prove that Venice is more than conventionally popular destinations such as the Piazza San Marco and Ponte di Rialto. Other places including churches, piazza, *campo*, gardens, exhibition venues, and even normal streets are also attracting people and making them realize the beauty of the city with different focal points. Further visualizations, comparisons, and discussions of the spatial mapping of OUV and HA categories identified with the proposed methodology can be found in Appendix B with Figs. B.4 till B.7.

5. Discussion

5.1. Documenting knowledge for heritage studies

The initial motivation for conducting this research is to propose a "knowledge documentation and mapping tool of cultural and natural heritage characteristics", especially the heritage values and attributes, for the "recognition of cultural significance and diversity", in support of the HUL approach (UNESCO, 2011). Instead of actively engaging the civil society to contribute to the narratives with their knowledge and values a city they live in or visit conveys to them, this study makes use of the existing information on social media with a real-world dataset to make exploratory analyses. The term "exploratory" is crucial for interpreting the findings and applying the methodology in practice. It functions as a complementary tool to help heritage managers and authorities explore the voices of the public on social media, either to confirm or to challenge/ adjust their hypotheses over the spatial distribution of the cultural significance in a city. For example, one could be affirmative ahead of time that tourists are over-crowded in only a few popular spots in Venice such as San Marco and Rialto, and that the beauties hidden in the other places are easily over-looked. However, the mapping practice in Fig. 12 suggests that Flickr users are indeed exploring a broad range of places all over the island, attracted by different types of cultural significance reflecting various heritage

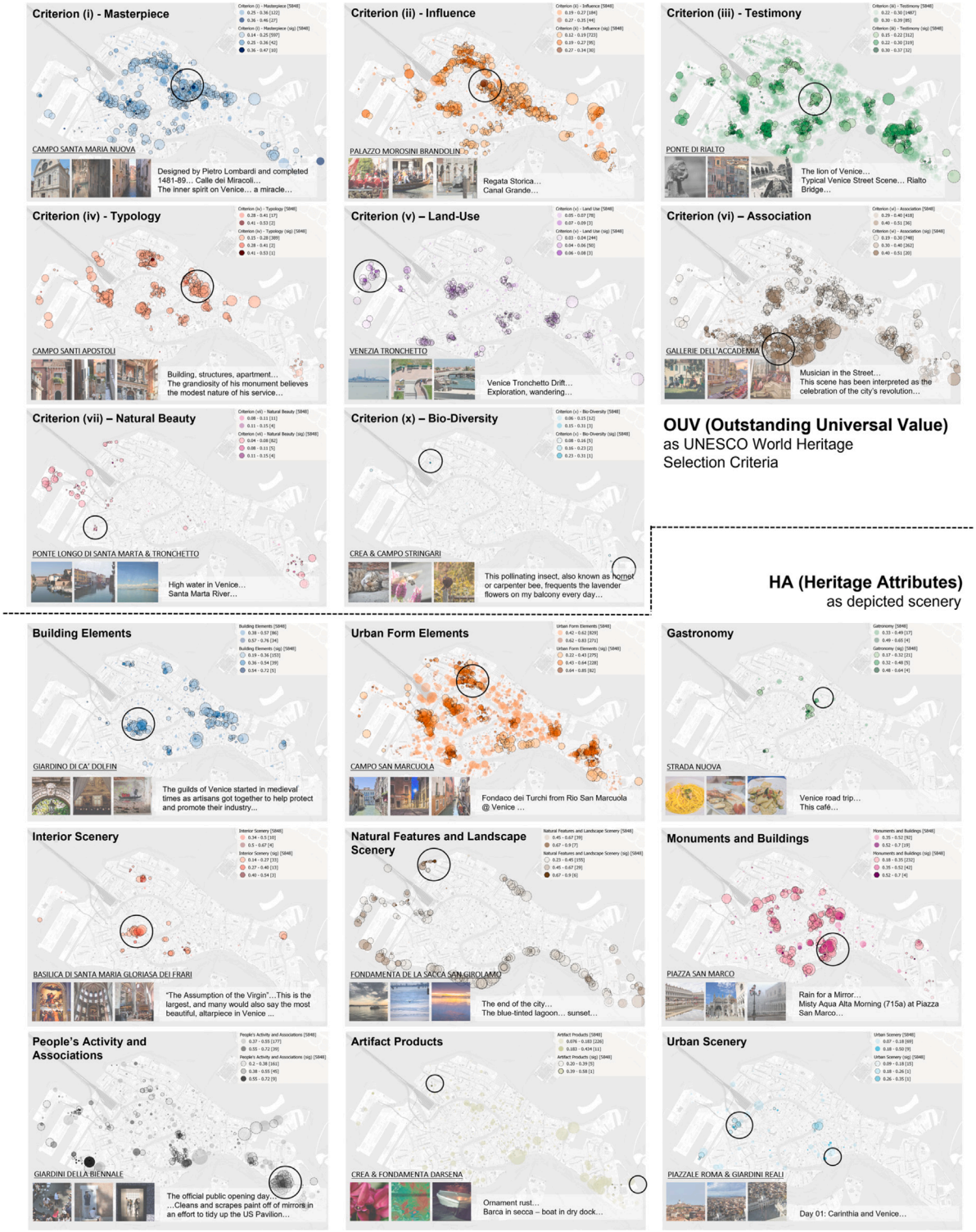

Fig. 12. The geographical distribution of OUV and HA categories in *VEN-XL* based on the spatial diffusion of labels. The nodes with high ranges of value for each category under equal-interval division are visualized as circles, the size of which demonstrates the number of posts distributed near the spatial node, while those nodes with a significant local Moran's *I* are shown with dashed borders. Three demonstrative photos and one comment from "hotspot" areas of categories are given below each map.

values and attributes. Heritage experts and practitioners could inspect the social media posts located nearby unexpected places revealed with cultural significance to get inspiration for further planning actions in pursuit of social inclusion (Waterton et al., 2006; Bai et al., 2021b).

In order to fully reflect the need for inclusive heritage management processes, further studies are needed to: (1) quantitatively and qualitatively collect ideas from broader communities, especially from those who do not use social media, for a fair comparison to justify the representativeness of similar studies; (2) apply the same methodology and test the models in a wider selection of case studies in different geographical and topological contexts, as to evaluate the generalizability of the proposed workflow; (3) update the OUV selection criteria and Heritage Attributes label categories with other frameworks, tailor-made for the research interests and objectives in their own usage scenarios. Furthermore, UNESCO Statements of OUV are assumed to include elements from both heritage values and attributes. This study completes one side of the puzzle of analysing the association between OUV Selection Criteria and Heritage Attributes and further mapping them spatially. Future studies could complete the other end by employing analyses and mapping practices under the classification framework of Heritage Values (Pereira Roders, 2007; Tarrafa Silva and Pereira Roders, 2010; Foroughi et al., 2022).

5.2. A mapping tool for urban explorations

Nevertheless, as a mapping tool in full mathematical details, the application scenarios of this study could go beyond heritage studies. In principle, given a back-end spatial network, the mathematical constructs of attention-based information aggregation and graph diffusion processes described in Section 2.3.4 could also be fed with any sort of input feature array obtained from posts instead of only the output labels to be aggregated and mapped on spatial nodes. For example, one could map the *SUN* attribute feature of "biking" or "socializing" to explore the activities distributed in a city or map the number and proportion of faces in the posted images to observe the crowdedness, or even map some low-level visual features to mine the patterns of architectural style (Sun et al., 2022). In this sense, the proposed methodology could be generalized in applications of measuring safety (by diffusing crime rate), vitality (by mapping diversity of human activity), and popularity of urban spaces (by plotting the crowdedness), where it diffuses any sort of human-generated information onto a spatial network with inherent connectivity patterns. It is clearly related to the location-led place profiling approach in Lai (2019), whereas the categories in this study go beyond the text-only clustering of urban activities.

When making spatial statistical inferences, like other similar spatial analyses, the result is dependent on how the spatial connectivity and weights are measured. An interesting alternative could be aggregating the posts on regular spatial grids of different resolutions and using queen/rook-based contiguity as weight matrix to perform the diffusion (Anselin, 2003; Rogerson, 2021). As such, the label information will be rasterized and can be easily overlayed and collated in GIS platforms with other global and local datasets (Esch et al., 2017; Bekker, 2020). Moreover, the diffusion-mapping process proposed by this paper can be seen as an alternative and supplement to the conventional kernel-density heatmaps, which is further elaborated upon in Appendix B.

Even though there are originally three types of graph links in Heri-Graphs (Bai et al., 2022), this study only discovers the mapping, aggregation, and diffusion on the spatial-level nodes for pragmatic reasons, since spatial mapping is the most desired option. However, other than diffusing spatial-level node labels, mapping the foci and interests to temporal nodes (time periods in history) and social nodes (groups of social media users) that are derivable from $\boldsymbol{A}^{\mathrm{TEM}}, \boldsymbol{A}^{\mathrm{SOC}}$ can also answer interesting research questions. For instance, other than the spatial bipartite relation $\boldsymbol{B}$ mentioned in Section 2.3.4, the temporal bipartite relation $\boldsymbol{B}^{\mathrm{TEM}}$ (mapping the posts to the unique sorted weekly timestamps) and the tri-diagonal temporal adjacency matrix $\boldsymbol{W}^{\mathrm{TEM}}$ (recording the consecutive patterns of the weekly timestamps) can be used to substitute the aggregation computation in Eq. (11) and the diffusion computation in Eq. (16). Here a similar relationship also holds according to Bai et al. (2022): $\boldsymbol{A}^{\mathrm{TEM}} = \left(\boldsymbol{B}^{\mathrm{TEM}}\boldsymbol{W}^{\mathrm{TEM}}\boldsymbol{B}^{\mathrm{TEM}^{\mathsf{T}}} > 0\right) = \boldsymbol{B}^{\mathrm{TEM}}\left(\boldsymbol{W}^{\mathrm{TEM}} > 0\right)\boldsymbol{B}^{\mathrm{TEM}^{\mathsf{T}}} \in \{0,1\}^{K\times K}$. Every other module of the methodological framework visualized in Fig. 1 is still valid, except that the aggregation and diffusion would be conducted on the temporal-level graph. Analogue to the 2-dimensional mapping of spatial labels presented in this study, 1-dimensional mapping of temporal labels could result in attributed timelines showing the development of different label and/or feature categories. Similar mapping computations can be conducted for the social graph (social network of users on social media). These effects will be discovered in follow-up studies in various use cases.

5.3. A machine learning application

It is worth noting that the labels generated in *VEN* and *VEN-XL* datasets were originally not annotated by humans, but rather by a few ML models, or more specifically, MLP models as connectors between hidden features and output soft-label vectors (Bai et al., 2022). Therefore, using more complex graph-based GNN models in this study to replicate labels generated by simple MLP seems a reversed knowledge distillation process (i.e., confident students teaching a group of teachers) (Gou et al., 2021). It has also been shown in the most recent literature that simple MLPs using a Bag of Words could outperform most graph-based models in text classification tasks (Galke and Scherp, 2022). This trend is again visible here for some of the metrics in Tables 3 and 4. However, this paper also shows that GNN models have other benefits in terms of inductive learning and missing input data, as demonstrated in Fig. 6. Considering that the pseudo-labels of training and validation sets came from data-points of high prediction confidence (with high top-n prediction logits) and consistency (with similar prediction results by different trained models), the philosophy behind the training process in this paper also resembles the self-training strategy, where the originally unlabelled samples that end up with top prediction confidence in one round of training are added to the next round as labelled ones (Li et al., 2018; Sun et al., 2020; Wang et al., 2022b). The indications of such similarities mentioned above to the methodology and results are, however, out of the scope of this paper.

The classification performance can be further improved by adding humans in the loop with active learning (Prince, 2004). An important challenge given by the Heri-Graphs dataset that is not yet solved in this study is the imbalance of categories and the extreme sparsity in some small classes. This is a pragmatic difficulty since Heri-Graphs were originally created with real-world social media data for an application in heritage studies and did not enforce the categories to be balanced (Bai et al., 2022). However, future studies could implement data augmentation on the small classes in the unbalanced training data to further improve the classification performance. Few-shot learning and Zero-shot learning techniques can also be implemented (Sung et al., 2018). Further specific investigations are also invited to discover the effect of different graph structures, e.g., the original weighted adjacency matrices instead of binary ones, for the training and diffusion processes.

While applying the obtained model from this study to other case study cities in the world, such as Amsterdam and Suzhou also collected by Bai et al. (2022), two options could be considered, following the conventional GNN terminology of *transductive* and *inductive* learning (Kipf and Welling, 2016; Yang et al., 2016; Hamilton et al., 2017; Velickovic et al., 2017). By stacking the graphs of different datasets together before sampling sub-graphs, the pre-trained models could be used to fine-tune the new models while the test data could be seen together with training data, entailing a *transductive learning* setting. On the contrary, directly applying the trained model here to other cases would mean that the new test data are totally unobserved during training,

entailing an *inductive learning* setting. Researchers are welcome to explore the advantages and drawbacks of either option according to their own application scenarios.

5.4. Related works about the workflow

The proposed workflow in Fig. 1 takes inspiration from many different fields.

The first main component, i.e., semi-supervised learning of multiple models (Section 2.3.2), was the initial motivation of Graph Neural Networks (Kipf and Welling, 2016) and has been a topic extensively studied in computer science, with or without a graph structure (Blum and Mitchell, 1998; Zhou and Li, 2010; Yang et al., 2016; Hamilton et al., 2017; Velickovic et al., 2017; Li et al., 2018; Ma and Tang, 2021). The extra complexity of this study from a real-world dataset is that the semi-supervised learning process needs to react to two modalities (visual and textual, among which the textual features might be missing) and perform well in two classification tasks (OUV and HA) with a multi-graph structure (composed of spatial, temporal, and social links). The most closely relevant study in the literature is Liu and De Sabbata (2021), which did not include the other two components, as already mentioned in Section 1.

The second main component, i.e., aggregating model predictions (Section 2.3.3), leverages the concept from Ensemble Learning (Schapire and Singer, 1998; Zhou, 2012; Sagi and Rokach, 2018). The approach of computing an aggregated prediction vector as a weighted average of multiple models is similar to the "soft voting" mechanism (Zhou, 2012). Outside the field of computer science, aggregating the opinions of multiple actors based on their agreement and confidence is also an active topic in decision science (Stone, 1961; Budescu and Rantilla, 2000; Budescu and Yu, 2007). However, it is a technical innovation in this study to assign a class-level agreement vector to each aggregated prediction by computing SVD on the matrices composed of the original predictions of models in the ensemble, which is informative for evaluating the effect of aggregation.

The third main component, i.e., aggregating and diffusing post-level labels onto spatial graphs (Section 2.3.4), contains the most methodological innovations of the proposed workflow. As already pointed out in Section 2.3.4, the processes of aggregating and diffusing information on graphs resemble the operations of graph pooling and graph filtering, respectively (Ma and Tang, 2021), thus the Eqs. (10) and (12) can be formally similar to the ones in Graph Neural Network literature (Velickovic et al., 2017; Lee et al., 2019; Knyazev et al., 2019). However, they are for different purposes: instead of computing intermediate representations for the training loop, in this paper, these Equations are used to summarize the post-level information and assign it to spatial nodes, which were initially unlabelled in nature. The exchange of label information on bipartite graphs as shown in Eq. (11) also makes it different from the Label Propagation Algorithm (Zhu and Ghahramani, 2002; Huang et al., 2020; Wang and Leskovec, 2021), albeit the latter approach has the same spirit of diffusing soft labels based on the connectivity of nodes. Even though plenty of studies attempted to draw the label categories of social media posts on spatial maps, the majority of them either directly plotted the posts as unconnected data points (Huang et al., 2019; Liu and De Sabbata, 2021), or provided only the predominant categories or word-clouds for each detected/predefined cluster (Hu et al., 2015; Lai et al., 2017; Ginzarly et al., 2019), or created a kernel-density heatmap to show the distribution without a mathematical expression for the spatial nodes (Lansley and Longley, 2016; Bekker, 2020; Kang et al., 2021). The proposed method has the benefit of keeping a soft label structure (as probability distribution) for each discrete spatial unit (street intersections), which is also algebraically derivable. Further advantages of the proposed mapping process with label diffusion will be elaborated with Figs. B.4, B.5, and associative discussions in Appendix B.

Interestingly, even though the process of aggregating and diffusing labels is rare in spatial mapping, an essentially similar approach can be found on social networks for developing recommendation systems, where information is diffused on a tripartite graph of user-image-tag (Mao et al., 2016; Zhang et al., 2017; Wang et al., 2018), which could be regarded an analogue of the space-post-label triplet in this study. Furthermore, an interesting connection can also be found in a few recent studies with label diffusion processes during semantic segmentation on point clouds (Mascaro et al., 2021; Deng et al., 2022; Liao et al., 2022) and in a study predicting the effect of drug-disease association using diffusion on a bipartite graph (Xie et al., 2021).

Despite all the resemblances mentioned above, an additional innovation in this study is to bring all the components from different fields together in a holistic workflow and adapt them accordingly to solve a real-world research problem: mapping cultural significance categories obtained from social media platforms. To the best of the authors' knowledge, this study is the first to combine all these aspects with interdisciplinary knowledge, especially as the label category of interest is a unique example from the field of heritage studies, dominated by expert-based qualitative approach.

6. Conclusions

This paper proposes a workflow to obtain social perception maps concerning the cultural significance of places located in an urban spatial network using social media information. Several graph neural network models are trained with semi-supervised learning on attributed graph datasets with visual and textual nodal features of user-generated posts, effective on various evaluation metrics. The predicted post-level soft labels are aggregated considering the confidence and agreement of models, which are further aggregated and diffused on a back-end spatial network to obtain spatial-level labels. The distributions of spatial labels on heritage-related cultural significance categories are tested with spatial statistics and mapped with examples. The entire workflow is mathematically explained in detail and tested with the case study of Venice, shown to provide reasonable maps of cultural significance. The workflow can also be applied to other cities worldwide as a knowledge documentation tool collecting the voices of communities posting on the internet, with the ultimate goal of promoting socially inclusive heritage management processes, as suggested by the UNESCO Historic Urban Landscape approach. Moreover, the proposed methodology of diffusing human-generated location-based information onto the spatial network also has the potential for broader use scenarios in different domains of urban studies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The presented study is within the framework of the Heriland-Consortium. HERILAND is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 813883. We are grateful for the constructive comments and inspiring suggestions provided by the reviewers and the editors. The discussions within the teams of HEVA, Heriland, and SpaceTimeLab are highly appreciated. Special thanks to Dr. Linde Egberts for her comments on an early draft of this paper.

Appendix A. Implementation details of models

For all models, Adam (Kingma and Ba, 2017) with L2 regularization of 2e-4 is used as the optimizer. The hyper-parameter tuning, model training, and inference on *VEN* are performed on NVIDIA GeForce RTX 3060 GPU, and the inference on *VEN-XL* is performed on Intel Core i7-12700KF CPU since it is too large to fit in GPU. Hyper-parameter tuning is performed in a small range with grid-search. The detail of training, the resource occupancy, and the inference time are given respectively in the following sections and in Table A.1. The training curves of the models for all four main evaluation metrics mentioned in Section 3.2.2 during training are visualized in Fig. A.1.

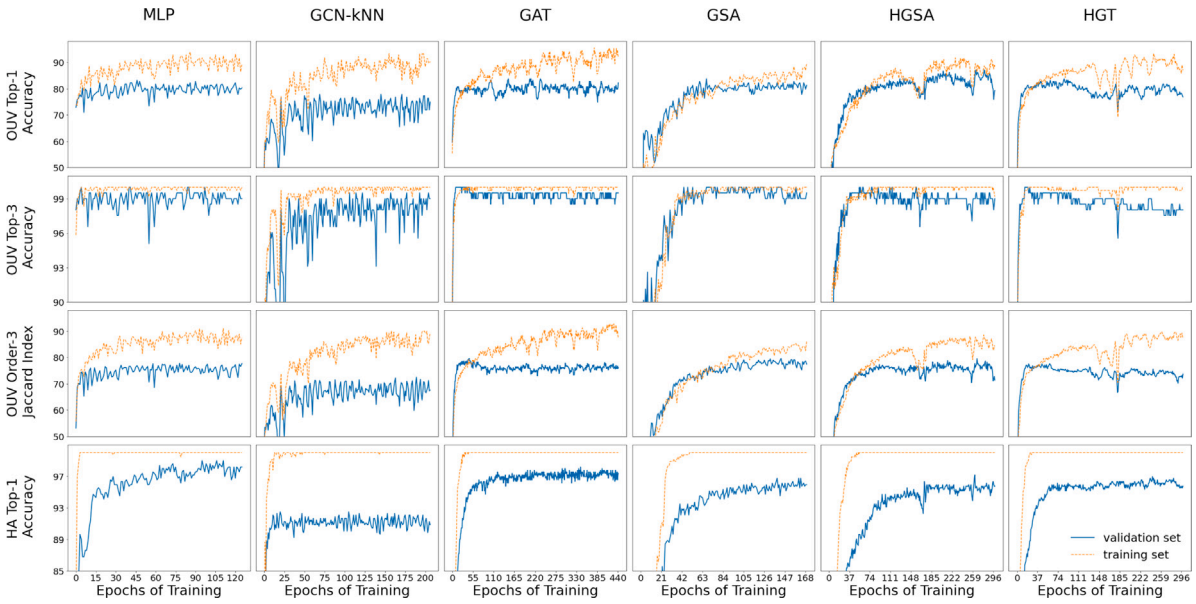

Fig. A.1. The training curves of the stored model checkpoints on the four main evaluation metrics for OUV and HA classification tasks. The dashed curves in orange show the performance of models on training set for each epoch, and the continuous curves in blue show the performance on validation set. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table A.1
The training resource occupancy, the model checkpoint size, and inference time (per each mini-batch) of each type of models.

Model	Number of epochs at early-stopping	Model size	Training time	Inference time GPU (*VEN*)	Inference time CPU (*VEN-XL*)
MLP	126/300	2.1 MB	0.02 s	0.02 s	0.33 s
GCN-kNN	207/500	115.2 MB	0.02 s	0.01 s	0.05 s
GAT	442/1000	6.0 MB	0.05 s	0.03 s	4.18 s
GSA	170/300	13.6 MB	0.09 s	0.06 s	13.54 s
HGSA	300/300	1.6 MB	0.03 s	0.03 s	3.39 s
HGT	300/300	0.6 MB	0.04 s	0.02 s	1.33 s

RDC No hyperparameter is tuned for the random classifier. The random choice function of Numpy library is used to generate top-3 OUV and top-1 HA predictions for each data sample based on the initial prior distribution of classes.

MLP The training takes 300 epochs with early-stopping criterion of 30 epochs. The hyper-parameters being tuned include learning rate in {.01, .001, .0005}, drop out rate in {.1, .2, .5}, number of hidden layers in {2, 3, 5}, and the size of hidden layers in {32, 64, 128, 256, 512}. The final selected model has a learning rate of .001, dropout rate of .1, and 3 hidden layers each with a size of 256.

GCN The training takes 500 epochs with early-stopping criterion of 100 epochs. The models use the initial residual connection alpha of 0.5, parameter to compute the strength of identity mapping theta of 1.0, and do not enable shared weights between the smoothed representation and the initial residuals. The hyper-parameters being tuned include learning rate in {.01, .001, .0001}, drop out rate in {.1, .2, .5}, number of hidden layers in {3, 6, 9}, and the size of hidden layers in {128, 256, 512, 1024, 2048}. The final selected model has a learning rate of .0001, dropout rate of .1, and 3 hidden layers each with a size of 2048. Furthermore, it turned out that the models using kNN links rather than the original graph structure perform better, therefore the same searched hyper-parameters are used to re-train a model checkpoint with kNN links as the final model.

GAT The training takes 1000 epochs with early-stopping criterion of 100 epochs. The models have two hidden GAT layers while the second one only has one attention head. The output of a linear hidden layer is concatenated with output of GAT filters before the final output layer. The hyper-parameters being tuned include learning rate in {.01, .001, .0001}, drop out rate in {.1, .3, .6}, number of attention heads for the first GAT layer in {2, 5, 8}, and the size of hidden layers in {32, 64, 128, 256, 512}. The final selected model has a learning rate of .0001, dropout rate of .1, 2 attention heads, and hidden layer size of 256.

GSA The training takes 300 epochs with early-stopping criterion of 30 epochs. The hyper-parameters being tuned include learning rate in {.01, .001, .0001}, drop out rate in {.1, .3, .5}, number of hidden layers in {2, 3, 5}, and the size of hidden layers in {32, 64, 128, 256, 512}. The final selected model has a learning rate of .0001, dropout rate of .1, and 5 hidden layers each with a size of 512.

HGSA The training takes 300 epochs with early-stopping criterion of 100 epochs. The output of a linear hidden layer is concatenated with output of Hetero GSA filters before the final output layer. The hyper-parameters being tuned include learning rate in {.01, .001, .0001}, number of hidden layers in {2, 3, 5}, and the size of hidden layers in {32, 64, 128, 256, 512}. The final selected model has a learning rate of .0001, and 3 hidden layers each with a size of 32.

HGT The training takes 300 epochs with early-stopping criterion of 100 epochs. The output of a linear hidden layer is concatenated with output of HGT before the final output layer. The hyper-parameters being tuned include learning rate in {.01, .001, .0005, .0001}, number of attention heads in {2, 4}, way of grouping attention heads in {sum, mean}, number of hidden layers in {2, 3, 5}, and the size of hidden layers in {32, 64, 128, 256}. The final selected model has a learning rate of .0005, 2 attention heads, grouping method of mean, and 3 hidden layers each with a size of 32.

Appendix B. Extended results

Table B.1 shows the post hoc comparison of the four different subsets $\mathcal{V}_{\text{train}}, \mathcal{V}_{\text{val}}, \mathcal{V}_{\text{test}}$, and $\mathcal{V}_{\text{unlab}}$ on their values of the confidence score $\boldsymbol{\kappa}^{\text{con}}$ and the agreement score $\boldsymbol{\kappa}^{\text{agr}}$, following the two-way ANOVA in Section 4.1. The difference between $\mathcal{V}_{\text{val}}$ and $\mathcal{V}_{\text{test}}$ is always insignificant, while all the other groups of comparisons have a significant difference with either moderate or very large effect sizes.

Table B.1
The post hoc comparison of the main effect of four different subsets for the confidence score $\boldsymbol{\kappa}^{\text{con}}$ and the agreement score $\boldsymbol{\kappa}^{\text{agr}}$ using the Tukey HSD Test.

Score	Group A	Group B	M(Group A)	M(Group B)	$\Delta(M)$	T	Tukey p	Cohen's d
Confidence score $\boldsymbol{\kappa}^{\text{con}}$	Training set $\mathcal{V}_{\text{train}}$	Validation set $\mathcal{V}_{\text{val}}$	0.746	0.663	0.083	89.315	<.0001	1.027
	Training set $\mathcal{V}_{\text{train}}$	Test set $\mathcal{V}_{\text{test}}$	0.746	0.664	0.082	88.638	<.0001	1.019
	Training set $\mathcal{V}_{\text{train}}$	Unlabelled set $\mathcal{V}_{\text{unlab}}$	0.746	0.564	0.182	210.015	<.0001	2.266
	Validation set $\mathcal{V}_{\text{val}}$	Test set $\mathcal{V}_{\text{test}}$	0.663	0.664	−0.001	−0.792	.858	−0.008
	Validation set $\mathcal{V}_{\text{val}}$	Unlabelled set $\mathcal{V}_{\text{unlab}}$	0.663	0.564	0.099	137.655	<.0001	1.239
	Test set $\mathcal{V}_{\text{val}}$	Unlabelled set $\mathcal{V}_{\text{unlab}}$	0.664	0.564	0.100	138.521	<.0001	1.247
Agreement score $\boldsymbol{\kappa}^{\text{agr}}$	Training set $\mathcal{V}_{\text{train}}$	Validation set $\mathcal{V}_{\text{val}}$	0.666	0.590	0.076	55.832	<.0001	0.642
	Training set $\mathcal{V}_{\text{train}}$	Test set $\mathcal{V}_{\text{test}}$	0.666	0.590	0.076	55.805	<.0001	0.642
	Training set $\mathcal{V}_{\text{train}}$	Unlabelled set $\mathcal{V}_{\text{unlab}}$	0.666	0.427	0.238	186.453	<.0001	2.012
	Validation set $\mathcal{V}_{\text{val}}$	Test set $\mathcal{V}_{\text{test}}$	0.590	0.590	0.000	−0.032	.999	−0.000
	Validation set $\mathcal{V}_{\text{val}}$	Unlabelled set $\mathcal{V}_{\text{unlab}}$	0.590	0.427	0.162	152.187	<.0001	1.370
	Test set $\mathcal{V}_{\text{val}}$	Unlabelled set $\mathcal{V}_{\text{unlab}}$	0.590	0.427	0.162	152.187	<.0001	1.370

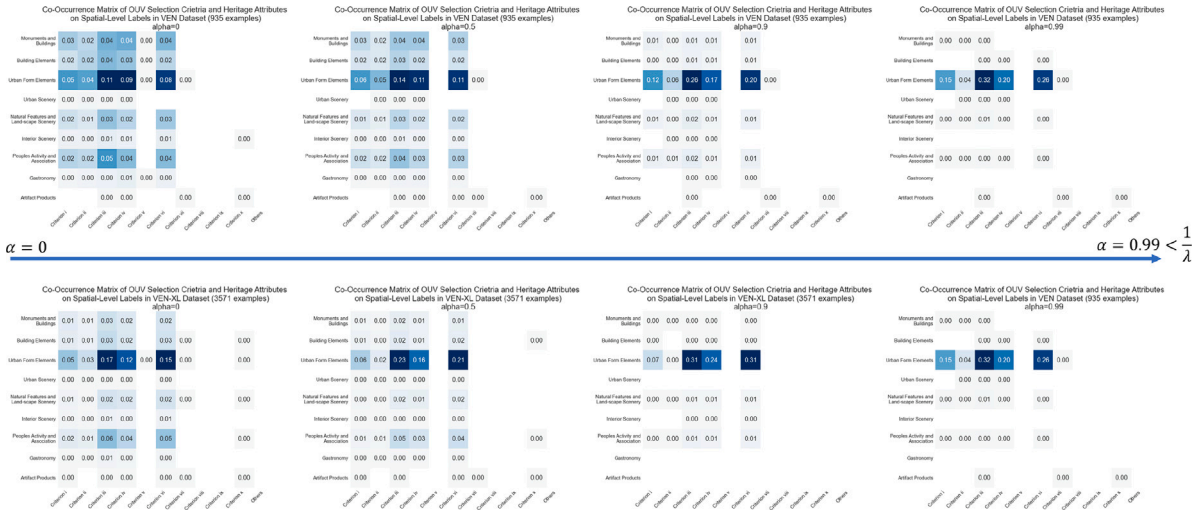

Fig. B.1. The change of normalized co-occurrence matrices $\boldsymbol{O}$ of the OUV and HA categories in spatial level label array $\boldsymbol{\mathcal{Y}}$ in both *VEN* and *VEN-XL* datasets, as the scaling parameter α changes.

Fig. B.1 shows the effect of α on the distribution of OUV and HA categories in the final diffused spatial label arrays $\boldsymbol{\mathcal{Y}}$. As α gets larger and closer to its theoretical maximum of $\min(1, 1/\lambda)$, the spatial labels get more to the extreme where all the labels are dominated only by the large classes. This is similar to the problem of "over-smoothing" in GNN literature (Li et al., 2018).

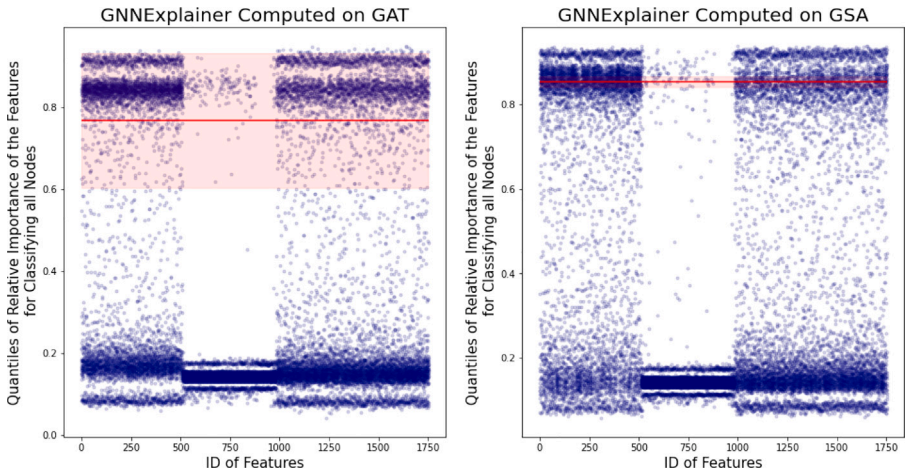

Fig. B.2. The scatter plots of all the 10-quantile values for the relative importance of all visual and textual features while classifying each node in $\mathcal{V}_{\text{train}}, \mathcal{V}_{\text{val}}, \mathcal{V}_{\text{test}}$ in GAT and GSA models, computed with GNNExplainer. The explainable visual features are with the indices of 512–981. The red lines and their shadows mark the means and standard deviations of the relative importance by the top-250_{th} feature.

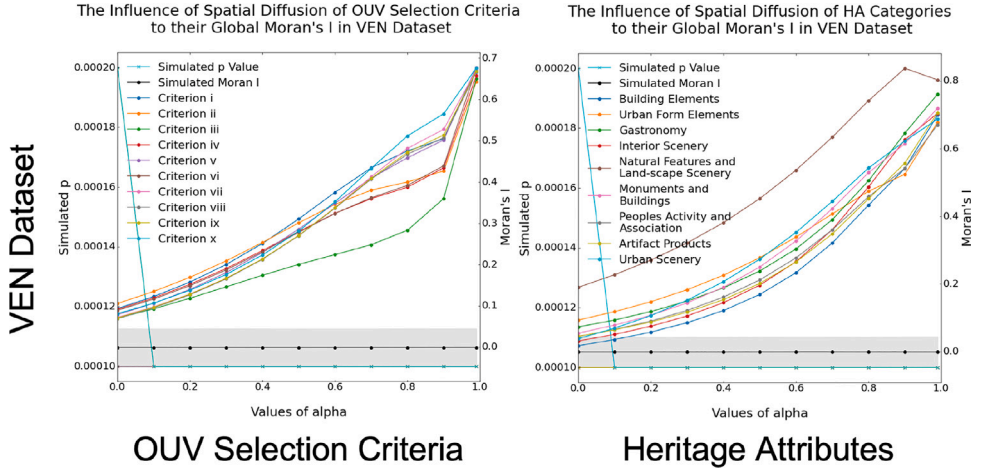

Fig. B.3. The change of global Moran's I in *VEN* with conventional row-standardized weight matrix only having zero diagonal entries. The Moran's I are generally smaller than in Fig. 10 since the self-correlations are not considered. For most categories, the spatial correlation is already significant without diffusion. For smaller α, the deviation of Moran's I is also smaller while significantly dropping the p values. Note the expected I value gets to the conventional scale of $-1/(N-1)$.

Computing the relative importance of all features while classifying each OUV/HA category using GNNExplainer will generate a soft mask vector for each node. Fig. B.2 plots all the 10-quantile values (similar to the median at the 50% partition, yet showing all values at the 10%, 20%, …, 90% partitions) of the soft mask values of each feature among all considered nodes, respectively using trained GAT and GSA as the base model. The distribution of the features shows that the relative importance computed by GNNExplainer on the explainable features is far less than that on the hidden features. How to explain and/or interpret those "non-explainable" hidden features would be an interesting future research direction. Inspecting the visualized distributions, that of GAT is slightly different from GSA in the sense that the hidden visual features (with the indices of 0–511, i.e., the left part of the images) are given higher relative importance in GSA. Furthermore, the red lines indicating the threshold of entering the top-250 entries for all the nodes imply that the two models work very differently using the information of all features. GAT has a lower top-250 threshold with a far wider confidence interval than GSA, suggesting that GAT uses very different numbers of features to predict the nodes, while the thresholds and thus the number of features being used in GSA are relatively more stable.

Fig. B.3 demonstrates a similar change pattern of Moran's I as in Fig. 10 with conventional definition of weight matrix:

$$I_C = \frac{(\boldsymbol{y}_C - \bar{y}_C\mathbf{1})^{\top}\tilde{\boldsymbol{W}}(\boldsymbol{y}_C - \bar{y}_C\mathbf{1})}{(\boldsymbol{y}_C - \bar{y}_C\mathbf{1})^{\top}(\boldsymbol{y}_C - \bar{y}_C\mathbf{1})}, \tag{B.1}$$

where the diagonal entries of $\tilde{\boldsymbol{W}}$ are all 0 and the row-sums of the matrix are all 1. Since a few spatial nodes in V (20 in *VEN* and 27 in *VEN-XL*) were isolated without any neighbours, rendering the row-standardization operation invalid, these nodes are omitted from the computation.

Figs. B.4 and B.5 respectively plot the distribution of high values on spatial nodes level for each OUV and HA category in *VEN* and *VEN-XL* datasets, and the high values on post levels overlapping with a kernel-density heatmap in *VEN* dataset only. A relatively stable pattern could be observed in the sense that the "hotspots" in *VEN* are generally detectable in *VEN-XL*, but not *vice versa*. In a few cases such as the HA category of Interior Scene, some significant clusters in *VEN* are diluted and no longer visible in *VEN-XL* with possibly more diverse post topics concerning OUV and HA. In general, the distribution in *VEN-XL* with more posts as data samples can be regarded as more reliable.

Fig. B.4. Comparison of the geographical distribution of post-level and spatial-level OUV node labels in *VEN* and spatial-level OUV labels in *VEN-XL* datasets. Post-level labels are accompanied by a kernel-density heatmap.

Note the methodology proposed in this study can be seen as an alternative and/or supplement to the conventional kernel-density heatmap weighted by the value in each channel. Figs. B.4 and B.5 also show the similarity and difference between the two methods in the case of *VEN* dataset. Generally, the hotspots are distributed in similar locations with both methods, since a spatial node can only be assigned high values when nearby posts also have high values consistently. However, the method proposed also considers confidence and agreement as crucial weighting parameters, preventing the risk in heatmaps that a very large number of medium-low values will also result in an overall hotspot in almost all categories, which is obvious in the case of San Marco square. Another benefit of the proposed method is that it is more specific and discretized than the kernel-density heatmap, yet more general and aggregated than mapping individual posts. The former is beneficial since it can point to certain places (street intersections) instead of only a broad region while tracing the posts as demonstrated in Fig. 12, easier for targeting useful information. The latter is beneficial since the method will not be too sensitive to individual posts while losing the main points. Furthermore, the proposed method performs aggregation on a fixed number of spatial nodes (a maximum of 5848 in Venice), easier for human comprehension, especially when the number of posts at hand grows to a larger scale, as demonstrated in Fig. B.6 where the top-right subplot mapping all the posts collected in Venice gets too crowded with points. However, Fig. B.6 also showcases another drawback of the dataset provided by Bai et al. (2022), that the spatial nodes only consisted of the ones on the main island and omitted places such as Giudecca island and San Giorgio Maggiore, pulling the posts on those places as well as on the canals to their nearest walkable spatial nodes on the southern harbour areas. This may have partially influenced the results of spatial distribution of categories such as OUV Criterion (vi) about Association and HA Natural Features and Landscape Scenery. This issue could be fixed in future studies by updating the assignment matrix $\boldsymbol{B}$ and spatial weight matrix $\boldsymbol{W}$.

Additionally, Fig. B.7 visualizes some typical posts of each OUV and HA category irrespective of their geographical locations, which can also be beneficial information for heritage scholars.

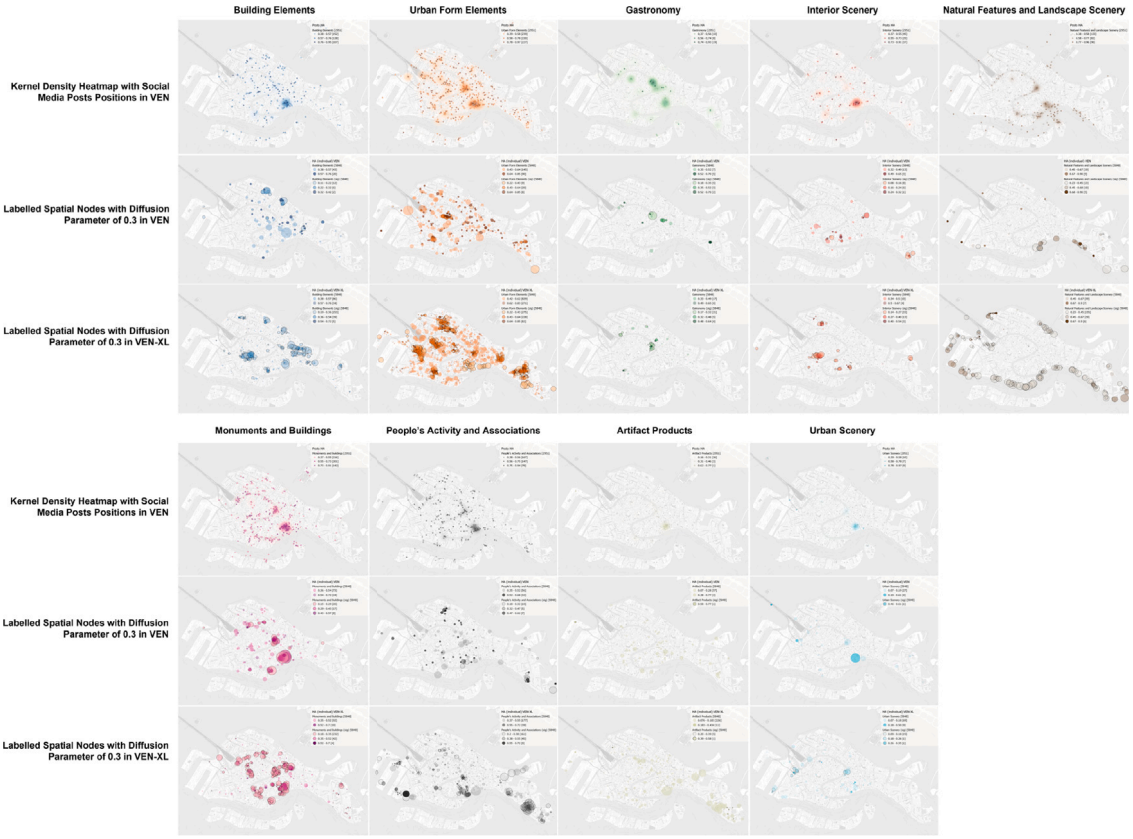

Fig. B.5. Comparison of the geographical distribution of post-level and spatial-level HA node labels in *VEN* and spatial-level HA labels in *VEN-XL* datasets. Post-level labels are accompanied by a kernel-density heatmap.

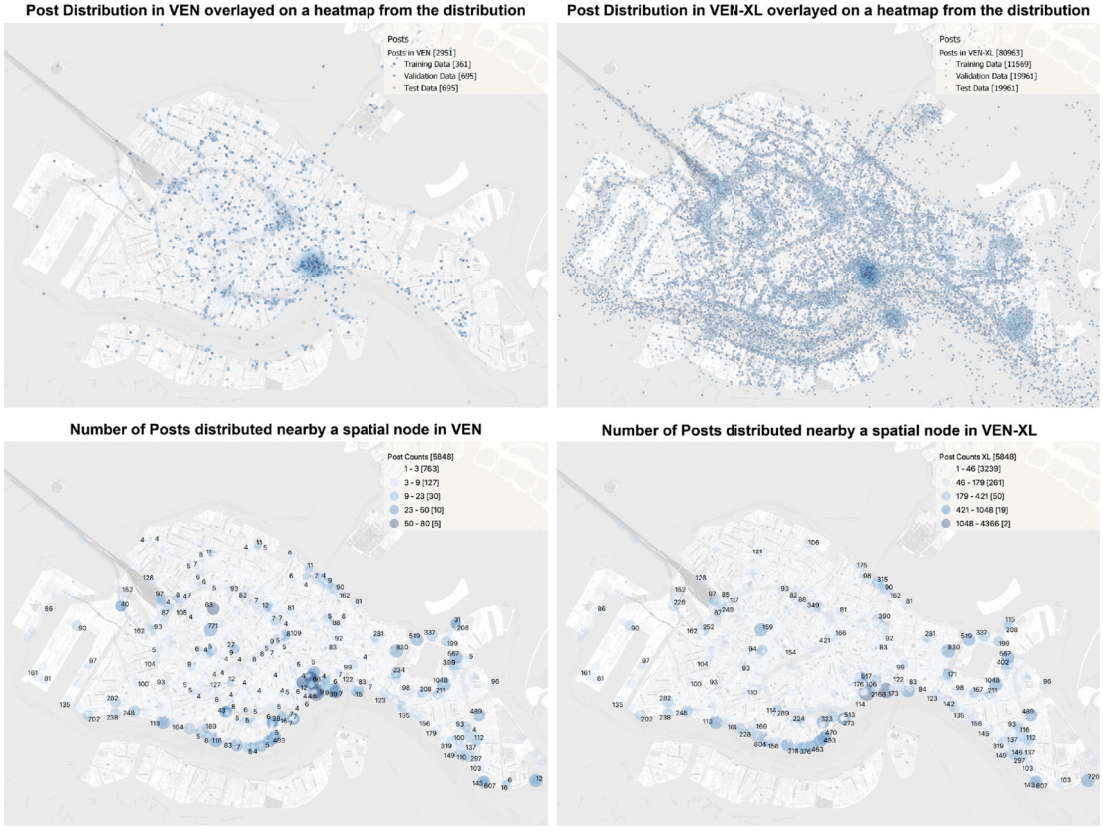

Fig. B.6. Top: the dis-aggregated distribution of all the geo-tagged posts in both *VEN* and *VEN-XL* datasets; Bottom: the number of posts distributed nearby each spatial node.

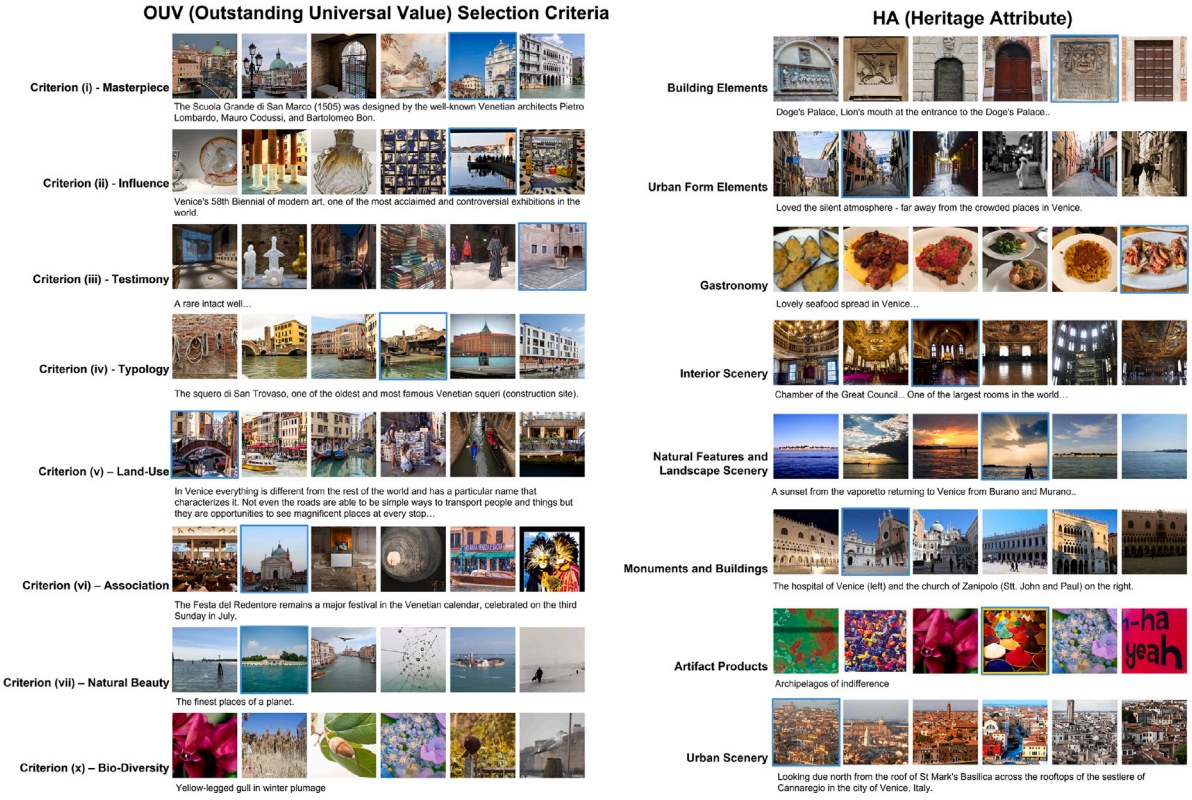

Fig. B.7. Post-level demonstrations of images and/or comments that have the largest logits for OUV and HA categories. For each category, six typical images and one comment are visualized, both are mostly among top-10 entries. The corresponding image to the comment is highlighted with a blue frame. No images from HA category People's Activities and Association are shown since the typical images always have a large portion of human faces on them. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Appendix C. Proof of equivalence

In this section, we will show that adding the last state of a node $\hat{\boldsymbol{y}}^{(t)}$ to the calculation of its current state during the diffusion process is equivalent to what has been proposed in Eqs. (12) and (13) for computing the steady-state $\boldsymbol{y}$.

Proof. By adding the term of the last state of a node itself, Eq. (12) could be adapted as:

$$\hat{\boldsymbol{y}}_k^{(t+1)} = \alpha_1 \hat{\boldsymbol{y}}_k^{(t)} + \alpha_2 \hat{\boldsymbol{y}}_k + \alpha_3 \frac{\sum_{v_{k'} \in \mathcal{N}_G(v_k)} W_{k,k'} \hat{\boldsymbol{y}}_{k'}^{(t)}}{\sum_{v_{k'} \in \mathcal{N}_G(v_k)} W_{k,k'}}, \tag{C.1}$$

or in its matrix form:

$$\hat{\boldsymbol{y}}^{(t+1)} = \alpha_1 \hat{\boldsymbol{y}}^{(t)} + \alpha_2 \hat{\boldsymbol{y}} + \alpha_3 \hat{\boldsymbol{y}}^{(t)} \left(\boldsymbol{W} \boldsymbol{D}^{-1} \right), \tag{C.2}$$

where $\alpha_1, \alpha_2, \alpha_2 \in [0, 1], \alpha_1 + \alpha_2 + \alpha_3 = 1$ are parameters balancing the importance of the last state of a node, the initial state of a node, and the last state of its neighbouring nodes. Then the steady state could be written as:

$$\boldsymbol{y} = \alpha_1 \boldsymbol{y} + \alpha_2 \hat{\boldsymbol{y}} + \alpha_3 \boldsymbol{y} \left(\boldsymbol{W} \boldsymbol{D}^{-1} \right), \tag{C.3}$$

$$\boldsymbol{y} \left((1 - \alpha_1) \boldsymbol{I} - \alpha_3 \boldsymbol{W} \boldsymbol{D}^{-1} \right) = \alpha_2 \hat{\boldsymbol{y}}, \tag{C.4}$$

$$\text{therefore, } \boldsymbol{y} = \frac{\alpha_2}{\alpha_2 + \alpha_3} \hat{\boldsymbol{y}} \left(\boldsymbol{I} - \frac{\alpha_3}{\alpha_2 + \alpha_3} \boldsymbol{W} \boldsymbol{D}^{-1} \right)^{-1}, \tag{C.5}$$

substituting the number $\alpha_3/(\alpha_2 + \alpha_3) \in (0, 1]$ with another parameter $\alpha_0 \in (0, 1]$, then Eq. (C.5) could be written as:

$$\boldsymbol{y} = (1 - \alpha_0) \hat{\boldsymbol{y}} \left(\boldsymbol{I} - \alpha_0 \boldsymbol{W} \boldsymbol{D}^{-1} \right)^{-1}, \tag{C.6}$$

exactly the same as Eq. (16). Here the parameter α_0 represents the relative importance of the last state of the neighbouring nodes of a node and its initial state, conceptually consistent with the original α mentioned in Section 2.3.4. □

It is worth noting that the diffusion chain presented here employs a Markov transition probability matrix but it is not a Markov Chain in its entirety because it is not a memory-less machine; in fact, the initial state contributes to the direction of the steady state vectors. Note by putting α_2 equal to zero we can turn this chain into a Markov Chain, in which case the $\boldsymbol{y}$ ends up being an eigenvector centrality array.

Appendix D. Nomenclature

In this section, all the mathematical notations used in this paper will be listed in Table D.1.

Table D.1
The nomenclature of mathematical notations used in this paper in alphabetic order.

Symbol	Data Type/Shape	Description
$\boldsymbol{A}$	Matrix of Boolean $\boldsymbol{A} := (\boldsymbol{A}^{\mathrm{TEM}} > 0) \bigvee (\boldsymbol{A}^{\mathrm{SPA}} > 0) \bigvee (\boldsymbol{A}^{\mathrm{SOC}} > 0) \in \{0,1\}^{K\times K}$	The adjacency matrix of all post nodes in the set $\mathcal{V}$ that have at least one link connecting them as a composed simple graph.
$\boldsymbol{A}^{(*)}$	Matrix of Boolean $\boldsymbol{A}^{(*)} := \left[A^{(*)}_{i,i'}\right]_{K\times K} \in \{0,1\}^{K\times K}$, $\boldsymbol{A}^{(*)} \in \{\boldsymbol{A}^{\mathrm{TEM}}, \boldsymbol{A}^{\mathrm{SPA}}, \boldsymbol{A}^{\mathrm{SOC}}\}$	The adjacency matrix of each of the three sub-graphs $\mathcal{G}^{(*)}$ of the multi-graph $\mathcal{G}$, "(*)" represents one of the link types in {TEM, SPA, SOC}.
$\boldsymbol{A}_s, \boldsymbol{A}_s^{(*)}$	Matrix of Boolean $\boldsymbol{A}_s, \boldsymbol{A}_s^{(*)} \in \{0,1\}^{\lvert\mathcal{V}_s\rvert\times\lvert\mathcal{V}_s\rvert}$	The sampled adjacency matrix in sub-graph $\mathcal{G}_s$ for model training and inference.
$\boldsymbol{A}^{\mathrm{kNN}}$	Matrix of Boolean $\boldsymbol{A}^{\mathrm{kNN}} := \left[A^{\mathrm{kNN}}_{i,i'}\right] \in \{0,1\}^{K\times K}$	The adjacency matrix of the k-Nearest Neighbour graph computed with visual features of posts.
$\alpha, \alpha_1, \alpha_2, \alpha_3$	Scalar Values $\alpha, \alpha_1, \alpha_2, \alpha_3 \in [0,1]$	The parameters adjusting the relative importance of neighbours in diffusion process.
$\mathcal{B}$	Bipartite Graph $\mathcal{B} = (\mathcal{V}, V, \mathscr{E}, \boldsymbol{B})$	The bipartite graph of postal nodes $\mathcal{V}$ and spatial nodes V with matrix $\boldsymbol{B}$ and edges $\mathscr{E}$.
$\boldsymbol{B}$	Matrix of Boolean $\boldsymbol{B} := \left[B_{i,k}\right] \in \{0,1\}^{K\times\lvert V\rvert}$	The bi-adjacency matrix of postal nodes $\mathcal{V}$ and spatial nodes V.
β	Scalar Value	The attenuation parameter for the computation of Katz centrality.
C	Integer Indices $C \in \{1,2,\dots,20\} \subset \mathbb{N}$	The index of the OUV and HA label category channels.
$\boldsymbol{D}$	Matrix of Floats $\boldsymbol{D} \in \mathbb{R}_+^{\lvert V\rvert\times\lvert V\rvert}$	A diagonal matrix where each entry records the weighted degree of graph G.
$\boldsymbol{e}_C$	1D Array of Boolean $\boldsymbol{e}_C \in \{0,1\}^{20\times 1}$	A one-hot unit vector marking the C_{th} entry as 1.
$\mathcal{F}$	A set of objects $\mathcal{F} = \{\mathbf{f}_j\}, j \in [0, \lvert\mathcal{F}\rvert)$	The set of candidate MLP or GNN models to be trained.
$\mathcal{G}$	Multi-Graph $\mathcal{G} = (\mathcal{V}, \{\mathcal{E}^{\mathrm{TEM}}, \mathcal{E}^{\mathrm{SPA}}, \mathcal{E}^{\mathrm{SOC}}\})$	The graph with temporal, spatial, and social links $\mathcal{E}^{(*)}$ among post nodes set $\mathcal{V}$.
$\mathcal{G}'$	Undirected Simple Graph $\mathcal{G}' = (\mathcal{V}, \mathcal{E})$	The simple composed graph of the multi-graph $\mathcal{G}$ with the same node set $\mathcal{V}$.
$\mathcal{G}_s$	Undirected Multi-Graph or Simple Graph, $\mathcal{G}_s = (\mathcal{V}_s, \{\mathcal{E}_s^{\mathrm{TEM}}, \mathcal{E}_s^{\mathrm{SPA}}, \mathcal{E}_s^{\mathrm{SOC}}\})$ or $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$	The sub-graphs sampled from the original graph $\mathcal{G}$ or $\mathcal{G}'$ to train the models and make inference.
G	Undirected Weighted Graph $G = (V, E, \boldsymbol{W})$	The backend geographical representation of the city as a spatial network.
γ, ϕ	Scalar parameters $\gamma, \phi \in \mathbb{R}$	The parameters to adjust the relative contribution of agreement and confidence scores in the computation of attention values $\boldsymbol{S}$.
i, i'	Integer Indices $i, i' \in \{0,1,2,\dots,K-1\} \subset \mathbb{N}$	The index of samples in the dataset.
I_C	Scalar Value of Float	The global Moran's I computed for the C_{th} label channel.
$\boldsymbol{I}_C$	1D Array of Float $\boldsymbol{I}_C \in \mathbb{R}^{\lvert V\rvert\times 1}$	The local Moran's I on all spatial nodes computed for the C_{th} label channel.
j	Integer Indices $j \in \{0,1,2,\dots,\lvert\mathcal{F}\rvert-1\} \subset \mathbb{N}$	The index of candidate models to be trained.
k, k'	Integer Indices $j \in \{0,1,2,\dots,\lvert V\rvert-1\} \subset \mathbb{N}$	The index of spatial nodes in the spatial network.
K	Integer	The sample size (number of posts).
$\boldsymbol{\kappa}^{\mathrm{con}}, \boldsymbol{\kappa}^{\mathrm{agr}}$	1D Array of Floats $\boldsymbol{\kappa}^{\mathrm{con}}, \boldsymbol{\kappa}^{\mathrm{agr}} \in [0,1]^{K\times 1}$	The prediction confidence and agreement value of the models in $\mathcal{F}$ for all the posts.
$\ell_{\mathrm{OUV}}, \ell_{\mathrm{HA}}$	Function returning Scalar Values	Topic-specific evaluation metrics for OUV and HA classification tasks.
$\mathcal{L}_{\mathrm{train}}, \mathcal{L}_{\mathrm{val}}^{\mathrm{V/A}}$	Function returning Scalar Values	The loss function of a training batch and the entire validation sets.
λ	Scalar Value	The largest eigenvalue of the matrix $\boldsymbol{W}\boldsymbol{D}^{-1}$
$\mathcal{N}_B, \mathcal{N}_G$	Function returning a set of nodes	The function returning the neighbours of a spatial node v_k in either the bipartite graph $\mathcal{B}$ as a set of postal nodes or the spatial network G as a set of spatial nodes.
$\omega_{V/A}$	Scalar parameter	The relative importance of OUV and HA performance during training.
$p_j, p_{*,j}^{\mathrm{V/A}(\cdot)}$	Scalar Values $p_j = p_{\mathrm{val,j}}^{\mathrm{OUV(nJ)}} + p_{\mathrm{val,j}}^{\mathrm{HA(1)}} + p_{\mathrm{test,j}}^{\mathrm{OUV(nJ)}} + p_{\mathrm{test,j}}^{\mathrm{HA(1)}} \in \mathbb{R}^+, p_{*,\mathrm{j}}^{\mathrm{OUV(1)}}, p_{*,\mathrm{j}}^{\mathrm{OUV(n)}}, p_{*,\mathrm{j}}^{\mathrm{OUV(nJ)}}, p_{*,\mathrm{j}}^{\mathrm{HA(1)}} \in [0,1]$	The value of a specific evaluation metric (top-1 accuracy, top-n accuracy, order-n Jaccard Index) in the validation or test set for OUV or HA categories by the model $\mathbf{f}_j$.
$\boldsymbol{s}_C$	1D Array of Floats $\boldsymbol{s}_C \in [0,1]^{K\times 1}$	The vector of attention values of all post nodes in $\mathcal{V}$ of the label channel C.
$\boldsymbol{S}$	2D Array of Floats $\boldsymbol{S} \in [0,1]^{20\times K}$	The matrix of attention values of all post nodes in $\mathcal{V}$ of all label channels.
$\sigma_{\boldsymbol{Z}_i,1}$	Scalar Value	The first singular value computed with SVD on the matrix $\boldsymbol{Z}_i$.
$\boldsymbol{\Theta}_j$	Array of Floats	The model parameter by the candidate model $\mathbf{f}_j$.
$\mathcal{V}$	A set of nodes $\mathcal{V} = \{v_i\}, i \in [0, K)$	The set of all nodes of posts in the graph $\mathcal{G}$.
$\mathcal{V}_{\mathrm{batch}}$	A set of nodes $\mathcal{V}_{\mathrm{batch}} \subset \mathcal{V}_{\mathrm{train}}, \mathcal{V}_{\mathrm{val}}, \mathcal{V}_{\mathrm{test}}$	The set of post nodes as mini-batches used for model training and inference.
$\mathcal{V}_{\mathrm{tex}\pm}$	A set of nodes $\mathcal{V}_{\mathrm{tex}+}, \mathcal{V}_{\mathrm{tex}-} \subset \mathcal{V}$	The set of post nodes with or without textual features.
$\mathcal{V}_{\mathrm{train}}, \mathcal{V}_{\mathrm{val}}, \mathcal{V}_{\mathrm{test}}, \mathcal{V}_{\mathrm{unlab}}$	A set of nodes $\mathcal{V}_{\mathrm{train}}, \mathcal{V}_{\mathrm{val}}, \mathcal{V}_{\mathrm{test}}, \mathcal{V}_{\mathrm{unlab}} \subset \mathcal{V}$	The set of post nodes respectively in the training set, validation set, test set, or unlabelled set.
$\mathcal{V}_{\mathrm{V}\pm,\mathrm{A}\pm}$	A set of nodes $\mathcal{V}_{\mathrm{V+,A+}}, \mathcal{V}_{\mathrm{V+,A-}}, \mathcal{V}_{\mathrm{V-,A+}}, \mathcal{V}_{\mathrm{V-,A-}} \subset \mathcal{V}$	The set of post nodes respectively with or without OUV or HA labels initially.
V	A set of nodes $V = \{v_k\}, k \in [0, \lvert V\rvert)$	The set of all spatial nodes of street intersections in the spatial network G.
$\boldsymbol{W}$	Matrix of Float $\boldsymbol{W} := [W_{k,k'}] \in [0,1]^{\lvert V\rvert\times\lvert V\rvert}$	The weighted adjacency matrix marking the temporal closeness of spatial nodes.
$\boldsymbol{X}$	2D Array of Floats $\boldsymbol{X} := \left[\boldsymbol{x}_i\right]_{i\in[0,K)} \in \mathbb{R}^{1753\times K}$	The visual and textual representation features of a post.
$\boldsymbol{X}_s$	2D Array of Floats $\boldsymbol{X}_s \in \mathbb{R}^{1753\times\lvert\mathcal{V}_s\rvert}$	The sampled input visual and textual features of nodes in sub-graph $\mathcal{G}_s$ used for model training and inference.
$\boldsymbol{X}^{\mathrm{tex}}$	2D Array of Floats $\boldsymbol{X}^{\mathrm{tex}} \in \mathbb{R}^{771\times K}$	The textual representation features of a post.
$\boldsymbol{X}^{\mathrm{vis}}$	2D Array of Floats $\boldsymbol{X}^{\mathrm{vis}} \in \mathbb{R}^{982\times K}$	The visual representation features of a post.
$\boldsymbol{y}_i^{\mathrm{HA}}, \boldsymbol{y}_i^{\mathrm{OUV}}$	1D Arrays of Floats $\boldsymbol{y}_i^{\mathrm{HA}} \in [0,1]^{9\times 1}, \boldsymbol{y}_i^{\mathrm{OUV}} \in [0,1]^{11\times 1}$	The HA and OUV labels of the node v_i if not empty
$\hat{\boldsymbol{y}}_{j,i}, \hat{\boldsymbol{y}}_{j,i}^{\mathrm{HA}}, \hat{\boldsymbol{y}}_{j,i}^{\mathrm{OUV}}$	1D Arrays of Floats $\hat{\boldsymbol{y}}_{j,i} \in [0,1]^{20\times 1} \hat{\boldsymbol{y}}_{j,i}^{\mathrm{HA}} \in [0,1]^{9\times 1}, \hat{\boldsymbol{y}}_{j,i}^{\mathrm{OUV}} \in [0,1]^{11\times 1}$	The predicted HA and OUV labels of the node v_i by the candidate model $\mathbf{f}_j$
$\hat{\boldsymbol{y}}_C$	1D Array of Floats $\hat{\boldsymbol{y}}_C := \hat{\boldsymbol{Y}}^{\mathsf{T}}\boldsymbol{e}_C \in [0,1]^{K\times 1}$	The labels of all post nodes in $\mathcal{V}$ for the C_{th} label channel.
$\boldsymbol{Y}_{\mathrm{V}\pm,\mathrm{A}\pm}$	2D Arrays of Floats or Empty Array	The "ground-truth" soft label arrays of post nodes respectively with or without OUV or HA labels initially.
$\hat{\boldsymbol{Y}}$	2D Array of Floats $\hat{\boldsymbol{Y}} := \left[\hat{\boldsymbol{y}}_i\right]_{v_i\in\mathcal{V}} \in [0,1]^{20\times K}$	The aggregated label array from $\hat{\boldsymbol{Y}}_j$ for all the posts by all the models in $\mathcal{F}$.
$\hat{\boldsymbol{Y}}_i$	2D Array of Floats $\hat{\boldsymbol{Y}}_i := \left[\hat{\boldsymbol{y}}_{j,i}\right]_{\mathbf{f}_j\in\mathcal{F}} \in [0,1]^{20\times\lvert\mathcal{F}\rvert}$	The predicted label array for the post v_i by all the models in $\mathcal{F}$.
$\hat{\boldsymbol{Y}}_j$	2D Array of Floats $\hat{\boldsymbol{Y}}_j := \left[\hat{\boldsymbol{y}}_{j,i}\right]_{v_i\in\mathcal{V}} \in [0,1]^{20\times K}$	The predicted label array for all the posts in $\mathcal{V}$ by the model $\mathbf{f}_j$.
$\hat{\boldsymbol{y}}_C$	1D Array of Floats $\hat{\boldsymbol{y}}_C := \hat{\boldsymbol{\mathcal{Y}}}^{\mathsf{T}}\boldsymbol{e}_C \in [0,1]^{\lvert V\rvert\times 1}$	The initial soft label value on all spatial nodes in the C_{th} label channel.
$\boldsymbol{y}_C$	1D Array of Floats $\boldsymbol{y}_C \in [0,1]^{\lvert V\rvert\times 1}$	The final soft label value on all spatial nodes in the C_{th} label channel after diffusion.
$\hat{\boldsymbol{\mathcal{Y}}}$	2D Array of Floats $\hat{\boldsymbol{\mathcal{Y}}} := \left[\hat{\boldsymbol{y}}_k\right] \in [0,1]^{20\times\lvert V\rvert}$	The aggregated spatial label array for spatial nodes from their nearby posts.
$\hat{\boldsymbol{\mathcal{Y}}}^{(t)}$	2D Array of Floats $\hat{\boldsymbol{\mathcal{Y}}}^{(t)} := \left[\hat{\boldsymbol{y}}_k^{(t)}\right] \in [0,1]^{20\times\lvert V\rvert}$	The diffused spatial label array for spatial nodes from their neighbours at the t_{th} iteration, where $\hat{\boldsymbol{\mathcal{Y}}}^{(0)} = \hat{\boldsymbol{\mathcal{Y}}}$.
$\boldsymbol{\mathcal{Y}}$	2D Array of Floats $\boldsymbol{\mathcal{Y}} := \left[\boldsymbol{y}_k\right] \in [0,1]^{20\times\lvert V\rvert}$	The diffused final spatial label array for spatial nodes from their spatial neighbours.
$\boldsymbol{z}_{j,i}^{\mathrm{HA}}, \boldsymbol{z}_{j,i}^{\mathrm{OUV}}$	1D Arrays of Floats $\boldsymbol{z}_{j,i}^{\mathrm{HA}} \in \mathbb{R}^{9\times 1}, \boldsymbol{z}_{j,i}^{\mathrm{OUV}} \in \mathbb{R}^{11\times 1}$	The hidden layer outputs by model $\mathbf{f}_j$ corresponding to HA and OUV label channels
$\boldsymbol{Z}_i$	Matrix of Floats $\boldsymbol{Z}_i \in [-1,1]^{20\times\lvert\mathcal{F}\rvert}$	The centred and normalized label matrix calculated from $\hat{\boldsymbol{Y}}_i$ for SVD computation.

References

Aggarwal, C.C., 2011. An introduction to social network data analytics. In: Aggarwal, C.C. (Ed.), Social Network Data Analytics. SPRINGER, pp. 1–15. http://dx.doi.org/10.1007/978-1-4419-8462-3_1, chapter 1.

Amato, F., Cozzolino, G., Di Martino, S., Mazzeo, A., Moscato, V., Picariello, A., Romano, S., Sperlí, G., 2016. Opinions analysis in social networks for cultural heritage applications. Smart Innov. Syst. Technol. 55, 577–586. http://dx.doi.org/10.1007/978-3-319-39345-2_51.

Anselin, L., 1995. Local indicators of spatial association—LISA. Geogr. Anal. 27 (2), 93–115.

Anselin, L., 2003. An Introduction to Spatial Autocorrelation Analysis with Geoda. Spatial Analysis Laboratory, University of Illinois, Champagne-Urbana, Illinois.

Bai, N., Luo, R., Nourian, P., Pereira Roders, A., 2021a. WHOSe heritage: Classification of UNESCO world heritage statements of "outstanding universal value" with soft labels. In: Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp. 366–384. http://dx.doi.org/10.18653/v1/2021.findings-emnlp.34.

Bai, N., Nourian, P., Luo, R., Pereira Roders, A., 2022. Heri-graphs: A dataset creation framework for multi-modal machine learning on graphs of heritage values and attributes with social media. ISPRS Int. J. Geo-Inf. 11 (9), http://dx.doi.org/10.3390/ijgi11090469.

Bai, N., Nourian, P., Pereira Roders, A., 2021b. Global citizens and world heritage: social inclusion of online communities in heritage planning. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLVI-M-1-2021, 23–30. http://dx.doi.org/10.5194/isprs-archives-XLVI-M-1-2021-23-2021.

Baltrusaitis, T., Ahuja, C., Morency, L.P., 2019. Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. 41 (2), 423–443. http://dx.doi.org/10.1109/TPAMI.2018.2798607, arXiv:1705.09406.

Bandarin, F., Van Oers, R., 2012. The Historic Urban Landscape: Managing Heritage in an Urban Century. John Wiley & Sons.

Bastian, M., Heymann, S., Jacomy, M., 2009. Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 3. pp. 361–362.

Baumer, E., Elovic, E., Qin, Y., Polletta, F., Gay, G., 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1472–1482.

Bekker, R., 2020. Creating Insights in Tourism with Flickr Photography, Visualizing and Analysing Spatial and Temporal Patterns in Venice (Master's thesis). Rijksuniversiteit Groningen.

Benzi, M., Klymko, C., 2014. A matrix analysis of different centrality measures. arXiv preprint arXiv:1312.6722.

Bertocchi, D., Visentin, F., 2019. "The overwhelmed city": Physical and social over-capacities of global tourism in venice. Sustainability 11 (24), 6937.

Bigne, E., Ruiz, C., Cuenca, A., Perez, C., Garcia, A., 2021. What drives the helpfulness of online reviews? A deep learning study of sentiment analysis, pictorial content and reviewer expertise for mature destinations. J. Destination Mark. Manag. 20, 100570.

Blum, A., Mitchell, T., 1998. Combining labeled and unlabeled data with co-training. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory. pp. 92–100.

Boeing, G., 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Comput. Environ. Urban Syst. 65, 126–139.

Bonacich, P., 1972. Factoring and weighting approaches to status scores and clique identification. J. Math. Sociol. 2 (1), 113–120.

Boy, J.D., Uitermark, J., 2017. Reassembling the city through instagram. Trans. Inst. Br. Geogr. 42 (4), 612–624.

Budescu, D.V., Rantilla, A.K., 2000. Confidence in aggregation of expert opinions. Acta Psychol. 104 (3), 371–398.

Budescu, D.V., Yu, H.-T., 2007. Aggregation of opinions based on correlated cues and advisors. J. Behav. Decis. Mak. 20 (2), 153–177.

Calvino, I., 1978. Invisible Cities. Houghton Mifflin Harcourt.

Cao, R., Tu, W., Yang, C., Li, Q., Liu, J., Zhu, J., Zhang, Q., Li, Q., Qiu, G., 2020. Deep learning-based remote and social sensing data fusion for urban region function recognition. ISPRS J. Photogramm. Remote Sens. 163, 82–97.

Cartwright, W.E., 2010. Addressing the value of art in cartographic communication. ISPRS J. Photogramm. Remote Sens. 65 (3), 294–299.

Chen, Y., 2021. An analytical process of spatial autocorrelation functions based on Moran's index. PLoS One 16 (4), e0249589.

Chen, M., Wei, Z., Huang, Z., Ding, B., Li, Y., 2020. Simple and deep graph convolutional networks. In: International Conference on Machine Learning. PMLR, pp. 1725–1735.

Cheng, T., Wicks, T., 2014. Event detection using Twitter: A spatio-temporal approach. PLoS One 9 (6), e97807.

Cho, N., Kang, Y., Yoon, J., Park, S., Kim, J., 2022. Classifying tourists' photos and exploring tourism destination image using a deep learning model. J. Qual. Assur. Hosp. Tour. 1–29.

Cosgrove, D., 1982. The myth and the stones of venice: an historical geography of a symbolic landscape. J. Hist. Geogr. 8 (2), 145–169.

Crandall, D., Backstrom, L., Huttenlocher, D., Kleinberg, J., 2009. Mapping the world's photos. In: WWW'09 - Proceedings of the 18th International World Wide Web Conference. pp. 761–770. http://dx.doi.org/10.1145/1526709.1526812.

Deng, Y., Wang, M., Yang, Y., Yue, Y., 2022. Hd-ccsom: Hierarchical and dense collaborative continuous semantic occupancy mapping through label diffusion. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 2417–2422.

Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., Zeidler, J., Dech, S., Strano, E., 2017. Breaking new ground in mapping human settlements from space–the global urban footprint. ISPRS J. Photogramm. Remote Sens. 134, 30–42.

Fey, M., Lenssen, J.E., 2019. Fast graph representation learning with PyTorch geometric. arXiv preprint arXiv:1903.02428.

Foroughi, M., de Andrade, B., Pereira Roders, A., 2022. Peoples' values and feelings matter: Participatory heritage management using social media. In: nola, J.M. (Ed.), Artificial Intelligence and Architectural Design, Vol. 33. pp. 107–120, chapter 6.

Galke, L., Scherp, A., 2022. Bag-of-words vs. Graph vs. Sequence in text classification: Questioning the necessity of text-graphs and the surprising strength of a wide MLP. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp. 4038–4051. http://dx.doi.org/10.18653/v1/2022.acl-long.279.

Gardner, M.W., Dorling, S., 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos. Environ. 32 (14–15), 2627–2636.

GeoMatt22, 2020-12-10. Similarity metrics for more than two vectors?. Stack Exchange. URL: https://stats.stackexchange.com/q/239211. (version: 2020-12-10) (access date: 2022-08-31).

Ginzarly, M., Pereira Roders, A., Teller, J., 2019. Mapping historic urban landscape values through social media. J. Cult. Herit. 36, 1–11. http://dx.doi.org/10.1016/j.culher.2018.10.002.

Gomez, R., Gomez, L., Gibert, J., Karatzas, D., 2019. Learning from #barcelona instagram data what locals and tourists post about its neighbourhoods. In: Leal-Taixe L., R.S. (Ed.), Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). In: LNCS, vol. 11134, Springer Verlag, pp. 530–544. http://dx.doi.org/10.1007/978-3-030-11024-6_41.

Gou, J., Yu, B., Maybank, S.J., Tao, D., 2021. Knowledge distillation: A survey. Int. J. Comput. Vis. 129 (6), 1789–1819.

Gould, P.R., 1967. On the geographical interpretation of eigenvalues. Trans. Inst. Br. Geogr. 53–86.

Gustcoven, E., 2016. Attributes of World Heritage Cities, Sustainability by Management–A Comparative Study Between the World Heritage Cities Of Amsterdam, Edinburgh and Querétaro (Master's thesis). KU Leuven.

Hamilton, W., Ying, Z., Leskovec, J., 2017. Inductive representation learning on large graphs. In: Advances in Neural Information Processing Systems, Vol. 30.

He, Z., Deng, N., Li, X., Gu, H., 2022. How to "read" a destination from images? machine learning and network methods for dmos' image projection and photo evaluation. J. Travel Res. 61 (3), 597–619.

Hu, Z., Dong, Y., Wang, K., Sun, Y., 2020. Heterogeneous graph transformer. In: Proceedings of the Web Conference 2020. pp. 2704–2710.

Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., Prasad, S., 2015. Extracting and understanding urban areas of interest using geotagged photos. Comput. Environ. Urban Syst. 54, 240–254.

Huang, Q., He, H., Singh, A., Lim, S.-N., Benson, A.R., 2020. Combining label propagation and simple models out-performs graph neural networks. arXiv preprint arXiv:2010.13993.

Huang, W., Li, S., 2016. Understanding human activity patterns based on space-time-semantics. ISPRS J. Photogramm. Remote Sens. 121, 1–10.

Huang, X., Wang, C., Li, Z., Ning, H., 2019. A visual–textual fused approach to automated tagging of flood-related tweets during a flood event. Int. J. Digit. Earth 12 (11), 1248–1264.

ICOMOS, A., 2013. The Burra Charter: The Australia ICOMOS Charter for Places of Cultural Significance 2013. Australia ICOMOS Incorporated.

Jacomy, M., Venturini, T., Heymann, S., Bastian, M., 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One 9 (6), e98679.

Jokilehto, J., 2007. Aesthetics in the world heritage context. In: Values and Criteria in Heritage Conservation. Polistampa, pp. 183–194.

Jokilehto, J., 2008. What is OUV? Defining the Outstanding Universal Value of Cultural World Heritage Properties. Technical Report, ICOMOS, ICOMOS Berlin.

Kang, Y., Cho, N., Yoon, J., Park, S., Kim, J., 2021. Transfer learning of a deep learning model for exploring tourists' urban image using geotagged photos. ISPRS Int. J. Geo-Inf. 10 (3), 137.

Katz, L., 1953. A new status index derived from sociometric analysis. Psychometrika 18 (1), 39–43.

Kingma, D.P., Ba, J., 2017. Adam: A method for stochastic optimization. arXiv:1412.6980.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Knyazev, B., Taylor, G.W., Amer, M., 2019. Understanding attention and generalization in graph neural networks. In: Advances in Neural Information Processing Systems, Vol. 32.
Lai, J., 2019. Urban Place Profiling Using Geo-Referenced Social Media Data (Ph.D. thesis). UCL (University College London).
Lai, J., Cheng, T., Lansley, G., 2017. Improved targeted outdoor advertising based on geotagged social media data. Ann. GIS 23 (4), 237–250.
Lansley, G., Longley, P.A., 2016. The geography of Twitter topics in London. Comput. Environ. Urban Syst. 58, 85–96.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521 (7553), 436–444.
Lee, J., Lee, I., Kang, J., 2019. Self-attention graph pooling. In: International Conference on Machine Learning. PMLR, pp. 3734–3743.
Li, Q., Han, Z., Wu, X.-M., 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In: Thirty-Second AAAI Conference on Artificial Intelligence. pp. 1–8.
Li, Y., Tarlow, D., Brockschmidt, M., Zemel, R., 2015. Gated graph sequence neural networks. arXiv preprint arXiv:1511.05493.
Liao, L., Chen, W., Xiao, J., Wang, Z., Lin, C.-W., Satoh, S., 2022. Unsupervised foggy scene understanding via self spatial-temporal label diffusion. IEEE Trans. Image Process. 31, 3525–3540.
Liu, P., De Sabbata, S., 2021. A graph-based semi-supervised approach to classification learning in digital geographies. Comput. Environ. Urban Syst. 86, 101583.
Ma, Y., Tang, J., 2021. Deep Learning on Graphs. Cambridge University Press.
Mao, J., Lu, K., Li, G., Yi, M., 2016. Profiling users with tag networks in diffusion-based personalized recommendation. J. Inf. Sci. 42 (5), 711–722.
Mascaro, R., Teixeira, L., Chli, M., 2021. Diffuser: Multi-view 2d-to-3d label diffusion for semantic scene segmentation. In: 2021 IEEE International Conference on Robotics and Automation. ICRA, IEEE, pp. 13589–13595.
Monteiro, V., Henriques, R., Painho, M., Vaz, E., 2014. Sensing world heritage: an exploratory study of Twitter as a tool for assessing reputation. In: Murgante, B., Misra, S., Rocha, A., Torre, C., Rocha, J., Falcao, M., Taniar, D., Apduhan, B., Gervasi, O. (Eds.), Computational Science and Its Applications - ICCSA 2014, PT II. In: Lecture Notes in Computer Science, vol. 8580, Univ Minho; Univ Perugia; Univ Basilicata; Monash Univ; Kyushu Sangyo Univ; Assoc Portuguesa Investigacao Operac, pp. 404–419.
Moran, P.A., 1950. Notes on continuous stochastic phenomena. Biometrika 37 (1/2), 17–23.
Nourian, P., 2016. Configraphics: Graph Theoretical Methods for Design and Analysis of Spatial Configurations (Ph.D. thesis). TU Delft.
Nourian, P., Rezvani, S., Sariyildiz, I., van der Hoeven, F., 2016. Spectral modelling for spatial network analysis. In: Proceedings of the Symposium on Simulation for Architecture and Urban Design (SimAUD 2016). SimAUD, pp. 103–110.
Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The Pagerank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford InfoLab.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, Vol. 32.
Patterson, G., Hays, J., 2012. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2751–2758.
Patterson, G., Xu, C., Su, H., Hays, J., 2014. The SUN attribute database: Beyond categories for deeper scene understanding. Int. J. Comput. Vis. 108 (1), 59–81.
Pereira Roders, A., 2007. Re-Architecture: Lifespan Rehabilitation of Built Heritage (Ph.D. thesis). Technische Universiteit Eindhoven, http://dx.doi.org/10.6100/IR751759.
Pereira Roders, A., 2019. The Historic Urban Landscape approach in action: Eight years later. In: Reshaping Urban Conservation. Springer, pp. 21–54. http://dx.doi.org/10.1007/978-981-10-8887-2_2.
Prince, M., 2004. Does active learning work? A review of the research. J. Eng. Educ. 93 (3), 223–231.
Psarra, S., 2018. The Venice Variations: Tracing the Architectural Imagination. UCL Press.
QGIS Development Team, 2009. QGIS Geographic Information System. Open Source Geospatial Foundation, URL: http://qgis.osgeo.org.
Rey, S.J., Anselin, L., 2007. PySAL: A python library of spatial analytical methods. Rev. Reg. Stud. 37 (1), 5–27.
Rogerson, P.A., 2021. Spatial Statistical Methods for Geography. SAGE Publications Ltd.
Rogerson, P., Sun, Y., 2001. Spatial monitoring of geographic patterns: an application to crime analysis. Comput. Environ. Urban Syst. 25 (6), 539–556.
Rubinstein, R.Y., Kroese, D.P., 2013. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. Springer Science & Business Media.
Ruskin, J., 1879. The Stones of Venice. Crowell.
Ruskin, J., Quill, S., 2015. Ruskin's Venice: The Stones Revisited. Lund Humphries.
Sagi, O., Rokach, L., 2018. Ensemble learning: A survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 8 (4), e1249.
Schapire, R.E., Singer, Y., 1998. Improved boosting algorithms using confidence-rated predictions. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory. pp. 80–91.
Schlichtkrull, M., Kipf, T.N., Bloem, P., Berg, R.v.d., Titov, I., Welling, M., 2018. Modeling relational data with graph convolutional networks. In: European Semantic Web Conference. Springer, pp. 593–607.
Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 815–823.
Stone, M., 1961. The opinion pool. Ann. Math. Stat. 1339–1342.
Sun, K., Lin, Z., Zhu, Z., 2020. Multi-stage self-supervised learning for graph convolutional networks on graphs with few labeled nodes. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. pp. 5892–5899.
Sun, M., Zhang, F., Duarte, F., Ratti, C., 2022. Understanding architecture age and style through deep learning. Cities 128, 103787.
Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M., 2018. Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1199–1208.
Tarrafa Silva, A., Pereira Roders, A., 2010. The cultural significance of world heritage cities : Portugal as case study. In: Heritage and Sustainable Development. Évora, Portugal, pp. 255–263. http://dx.doi.org/10.13140/2.1.1152.0800.
Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. Econ. Geogr. 46 (sup1), 234–240.
UNESCO, 1972. Convention Concerning the Protection of the World Cultural and Natural Heritage. Technical Report november, UNESCO, Paris.
UNESCO, 2008. Operational Guidelines for the Implementation of the World Heritage Convention. Technical Report July, UNESCO World Heritage Centre.
UNESCO, 2011. Recommendation on the Historic Urban Landscape. Technical Report, UNESCO, Paris.
Urry, J., Larsen, J., 2011. The Tourist Gaze 3.0. SAGE Publications.
Vallat, R., 2018. Pingouin: statistics in python. J. Open Source Softw. 3 (31), 1026.
VanderWeele, T.J., Mathur, M.B., 2019. Some desirable properties of the Bonferroni correction: is the Bonferroni correction really so bad? Am. J. Epidemiol. 188 (3), 617–618.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 6000–6010.
Veldpaus, L., 2015. Historic Urban Landscapes: Framing the Integration of Urban and Heritage Planning in Multilevel Governance (Ph.D. thesis). Technische Universiteit Eindhoven, p. 210.
Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., et al., 2017. Graph attention networks. Stat 1050 (20), 10–48550.
Wang, L., Han, X., He, J., Jung, T., 2022a. Measuring residents' perceptions of city streets to inform better street planning through deep learning and space syntax. ISPRS J. Photogramm. Remote Sens. 190, 215–230.
Wang, Y., Jin, W., Derr, T., 2022b. Graph neural networks: Self-supervised learning. In: Wu, L., Cui, P., Pei, J., Zhao, L. (Eds.), Graph Neural Networks: Foundations, Frontiers, and Applications. Springer Singapore, Singapore, pp. 391–420.
Wang, H., Leskovec, J., 2021. Combining graph convolutional neural networks and label propagation. ACM Trans. Inf. Syst. (TOIS) 40 (4), 1–27.
Wang, P., Luo, H., Obaidat, M.S., Wu, T.-Y., 2018. The internet of things service recommendation based on tripartite graph with mass diffusion. In: 2018 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, pp. 1–6.
Waterton, E., Smith, L., Campbell, G., 2006. The utility of discourse analysis to heritage studies: The Burra Charter and social inclusion. Int. J. Herit. Stud. 12 (4), 339–355.
Wu, L., Cui, P., Pei, J., Zhao, L., 2022. Graph Neural Networks: Foundations, Frontiers, and Applications. Springer Singapore, Singapore, p. 725.
Xie, G., Li, J., Gu, G., Sun, Y., Lin, Z., Zhu, Y., Wang, W., 2021. BGMSDDA: a bipartite graph diffusion algorithm with multiple similarity integration for drug–disease association prediction. Mol. Omics 17 (6), 997–1011.
Xu, Y., Zhou, B., Jin, S., Xie, X., Chen, Z., Hu, S., He, N., 2022. A framework for urban land use classification by integrating the spatial context of points of interest and graph convolutional neural network method. Comput. Environ. Urban Syst. 95, 101807.
Yang, Z., Cohen, W., Salakhudinov, R., 2016. Revisiting semi-supervised learning with graph embeddings. In: International Conference on Machine Learning. PMLR, pp. 40–48.
Ying, Z., Bourgeois, D., You, J., Zitnik, M., Leskovec, J., 2019. Gnnexplainer: Generating explanations for graph neural networks. In: Advances in Neural Information Processing Systems, Vol. 32.
Yuster, R., Zwick, U., 2005. Fast sparse matrix multiplication. ACM Trans. Algorithms (TALG) 1 (1), 2–13.
Zancheti, S.M., Jokilehto, J., 1997. Values and urban conservation planning: some reflections on principles and definitions. J. Archit. Conserv. 3 (1), 37–51.
Zhan, J., Gurung, S., Parsa, S.P.K., 2017. Identification of top-k nodes in large networks using katz centrality. J. Big Data 4 (1), 1–19.
Zhang, Y., Cheng, T., 2020. Graph deep learning model for network-based predictive hotspot mapping of sparse spatio-temporal events. Comput. Environ. Urban Syst. 79, 101403.
Zhang, Y., Li, Y., Zhang, E., Long, Y., 2022a. Revealing virtual visiting preference: Differentiating virtual and physical space with massive TikTok records in Beijing. Cities 130, 103983.

Zhang, C., Song, D., Huang, C., Swami, A., Chawla, N.V., 2019a. Heterogeneous graph neural network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 793–803.

Zhang, J., Yang, Y., Tian, Q., Zhuo, L., Liu, X., 2017. Personalized social image recommendation method based on user-image-tag model. IEEE Trans. Multimed. 19 (11), 2439–2449.

Zhang, Y., Zhang, F., Chen, N., 2022b. Migratable urban street scene sensing method based on vision language pre-trained model. Int. J. Appl. Earth Obs. Geoinf. 113, 102989.

Zhang, F., Zhou, B., Ratti, C., Liu, Y., 2019b. Discovering place-informative scenes and objects using social media photos. R. Soc. Open Sci. 6 (3), 181375.

Zhou, Z.-H., 2012. Ensemble Methods: Foundations and Algorithms. CRC Press.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2017. Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40 (6), 1452–1464.

Zhou, Z.-H., Li, M., 2010. Semi-supervised learning by disagreement. Knowl. Inf. Syst. 24 (3), 415–439.

Zhu, X., Ghahramani, Z., 2002. Learning from Labeled and Unlabeled Data with Label Propagation. Tech. Rep., Technical Report CMU-CALD-02–107, Carnegie Mellon University.