Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-01505-4 OPEN

Explainable dimensionality reduction (XDR) to unbox AI ‘black box’ models: A study of AI perspectives on the ethnic styles of village dwellings

Xun Li[1]✉, Dongsheng Chen[2], Weipan Xu[1], Haohui Chen[3], Junjun Li[1] & Fan Mo[1]

Artificial intelligence (AI) has become frequently used in data and knowledge production in diverse domain studies. Scholars began to reflect on the plausibility of AI models that learn unexplained tacit knowledge, spawning the emerging research field, eXplainable AI (XAI). However, superior XAI approaches have yet to emerge that can explain the tacit knowledge acquired by AI models into human-understandable explicit knowledge. This paper proposes a novel eXplainable Dimensionality Reduction (XDR) framework, which aims to effectively translate the high-dimensional tacit knowledge learned by AI into explicit knowledge that is understandable to domain experts. We present a case study of recognizing the ethnic styles of village dwellings in Guangdong, China, via an AI model that can recognize the building footprints from satellite imagery. We find that the patio, size, length, direction and asymmetric shape of the village dwellings are the key to distinguish Canton, Hakka, Teochew or their mixed styles. The data-derived results, including key features, proximity relationships and geographical distribution of the styles are consistent with the findings of existing field studies. Moreover, an evidence of Hakka migration was also found in our results, complementing existing knowledge in architectural and historical geography. This proposed XDR framework can assist experts in diverse fields to further expand their domain knowledge.

[1] Sun Yat-sen University, Guangzhou, China. [2] Technical University of Munich, Munich, Germany. [3] Commonwealth Scientific and Industrial Research Organisation, Canberra, Australia. ✉email: lixun@mail.sysu.edu.cn

Introduction

The polymath Polanyi told us, “We could know more than we could tell”, implying the paradoxical fact that humans pursue “explicit” knowledge even though our knowledge is mostly “tacit”(Polanyi, 2009). Such paradox rejuvenates in the digital age where artificial intelligence (AI) and machine learning (ML) thrive.

The birth of eXplainable AI. The machine learns from human examples to derive the “tacit” knowledge about how to recognize objects (Russakovsky et al., 2015), process language (Devlin et al., 2018), and even drive vehicles (Grigorescu et al., 2020), being lack of pre-established rules gained from human’s explicit knowledge (Kambhampati, 2021). As the AI models continually deliver promising results, humans seem to accept the bitter fact that the machine’s decision process is mostly miserable and uninterpretable. However, scholars have started to reflect on such an uninterpretable decision process in recent years. Two main concerns about using machine’s tacit knowledge in real-world applications emerged from the ethical and technical perspectives, respectively. The former concerns whether the decisions are racially (Mehrabi et al., 2021; Angwin et al., 2022), gender-(Lu et al., 2020), or age-related biased (Díaz et al., 2018), while the latter concerns whether the machine “cheats” the learning system, e.g., in object recognition (Dombrowski et al., 2019; Lapuschkin et al., 2019), detecting melanoma (Winkler et al., 2019) and sentiment analysis and question answering (Wang et al., 2020).

For a long period, at least before the machine can develop explicit knowledge or rules from their tacit knowledge, humans would be exposed to the risk that the machine made a wrong judgment if their behaviour is not thoroughly audited. All the concerns discussed above reflect our human needs in pursuing explicit knowledge; that is, we need to know how the machine works before trusting them (Brundage et al., 2020). Kambhampati (Kambhampati, 2021) gave an interesting example that if an employee insists on learning how a company works purely based on observations and actions, refusing to learn operating procedures, the worker might be capable in some tasks but hardly a competent staff (e.g., failing to comply with company protocols). The process of understanding how AI works is referred to as an emerging research topic called eXplainable AI (XAI). XAI, as argued by both Miller (2019) and Lombrozo (2006), is both a process and a product from the social science perspective. While XAI identifies the causes for an event (e.g., why the AI models make such predictions with specific inputs), it also transfers knowledge between ‘explainer’ and ‘explainee’, enhancing human understanding of the corresponding domain knowledge (Roscher et al., 2020). For instance, using a neural network visualization tool such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) to visualize the toxic comment classifier, one can intuitively discover new toxic terms. Therefore, XAI allows humans to unpack machines’ tacit knowledge, enhancing explicit knowledge.

Domain knowledge in XAI. XAI is still a relatively new study area compared to the advances in AI algorithms in recent years. Various XAI methods are proposed in empirical studies, including dimension reduction, feature importance, attention mechanism, knowledge distillation and surrogate models (Yang et al., 2022). This study focuses on improving the dimension reduction method, as it’s universally applicable, enabling the basic behaviour of diverse AI models to be easily understood. Conventional dimension reduction benefits from unsupervised approaches, such as principal component analysis (PCA) (Wold et al., 1987; Islam et al., 2020a) and independent component analysis (ICA) (Comon, 1994). The high-dimension neural network features were transformed into human-friendly feature space of lower dimension. Afterward, domain experts tried to qualitatively interpret the latent meaning of each principal component by observing the corresponding activation patterns.

However, as the principal components are mixtures of the input neural network features, it is impossible to quantitatively translate the neural network features to the corresponding domain knowledge. Moreover, if the leading components (e.g., the first three components) account for large variances, different domain knowledge cannot be discriminated against each other in such a low dimension feature space. As a result, the unsupervised approach lacks a pathway for translating machine’s tacit knowledge into human explicit knowledge. Infusion of domain knowledge in dimension reduction emerges as an XAI method. The domain knowledge is summarized and translated into a supervised approach to explain how AI “black box” models work. Large amounts of empirical studies have proved such a method can improve the explainability and interpretability of AI models (Islam et al., 2020a, 2020b). However, few studies have proved that such a knowledge-infusion XAI method has generated new knowledge, which in turn benefits domain knowledge development. Today’s AI systems learn from millions of human examples so that they may observe hidden patterns in the data (Samek et al., 2017). The XAI method should allow extracting this ‘tacit’ knowledge into explicit knowledge. Therefore, this study aims to propose a novel dimension reduction framework for the infusion of domain knowledge, leading to better explainability of the AI models and the discovery of new domain knowledge. To be more specific, we summarize domain knowledge and build training samples (with human labels) to feed an XGBoost-based SHAP model to extract the most significant feature maps that are the outputs of a fine-tuned AI model. The framework is called eXplainable Dimensionality Reduction (XDR). More details can be found in the “Methodology” section.

A novel framework explaining typological characterization of rural dwellings. In this study, we use a remote sensing-based image segmentation model, Mask R-CNN, as the case study to demonstrate the effectiveness of the proposed framework. The segmentation model that was learned from more than 10,000 labels of the building footprints covering Guangdong province, China, has been proved to be effective in the segmentation of rural buildings with different layouts and styles (Li et al., 2021, 2022).

We focus on explaining how AI models recognize the ethnic style of rural village dwellings from the perspective of buildings’ typological characterization. Specifically, we try to understand whether the AI model could discriminate different building layouts and what spatial features help make such decisions. The prior architectural knowledge of historical geography in this context helps us to reinforce the understanding of how different neural network features affect the segmentation decisions.

Clustering historical buildings based on specific typological characterization have always been an integral part of human geography research, offering the understanding of the local ties between the ethnic inheritance, the territorial context, the natural environment, and agricultural activities. Over the past decades, scholars have accumulated flourishing architectural and cultural knowledge about historical buildings, allowing them to identify spatial clusters in many case studies in Europe (Fuentes et al., 2011; Ruggiero et al., 2019; Zanfi et al., 2020). In China, the historical and cultural knowledge of traditional villages is significant to the local community, promoting tourism (Gao and Wu, 2017) and building collective identity (Qin and Leung, 2021) (especially for the relatively poor remote areas). The typological characterization of Chinese traditional villages is a product of mass migrations (due to wars), conflicts between immigrants and aboriginal, and the local natural environment. Chinese scholars have collected and conceptualized historical buildings' characteristics through decades of field (Lu, 1981, 2007, 2008; Situ, 2001). However, to the best of our knowledge, few studies have systematically identified historical buildings and annotated the corresponding characteristics. Therefore, this study provides a novel XAI method to unpack a building segmentation model's convolutional features relating to typological characterization and using them as predictors to map different historical buildings at scale. The discussion about the case study area can be found in the section ”Study area and materials”.

In this work, our contributions are as follows:

1. We propose an XAI framework to infuse domain knowledge for dimension reduction of deep features of the AI black-box models, leading to better explainability and interpretability.
2. Dwellings' patio, size, length, direction and asymmetric shape are the key to distinguishing Canton, Hakka, Teochew or their mixed styles
3. Proximity relationships and geographical distribution of the styles are consistent with the findings of existing field studies.
4. Evidence of the fourth Hakka historical migration was also found.

Methods

Study area and materials. We use Guangdong Province, China as the case study area (see Fig. 1). As Guangdong had been the destination of large-scale internal migration and the origin of overseas out-migration, it is known for its cultural diversity. Moreover, the research community has accumulated and conceptualized lavish prior historical building knowledge for this area (Lu, 1981, 2008), so it facilitates the process of infusing domain knowledge for the proposed XAI framework.

In history, there were three waves of massive domestic migration due to the civil wars. As a result, Guangdong developed three independent ethnicities, including Canton, Hakka and Teochew (Table 1). The Canton ethnic group was formed during the Tang Dynasty. Its residence followed the traditional courtyard and *three-room* style in the central plains of north China at the very beginning. Additionally, livestock, kitchen storage, wells, and a three-in-one courtyard are attached to the three-room main building structure. The Teochew ethnic group was formed during the Tang and Song Dynasties, migrating southward along the coastline from southern Fujian. Its dwelling layout patterns include *Xiashanhu dwellings*, *Sidianjin dwellings*, and *Zhugancuo dwellings*. The Hakka ethnic group ultimately formed in the late Ming and Qing Dynasties and settled down in the mountainous areas of northern Guangdong with harsh natural conditions. The Hakka dwellings have the most distinctive types, including *Enclosed houses*, *Hakka Earth buildings* and *Circle dwellings* (Situ, 2001; Lu, 2007).

Two datasets were used in this study, the high-resolution satellite imageries and the place of interest (POI). The satellite imageries were derived through MapQuest (www.mapquest.com), a satellite imagery provider in the United States. It covers the

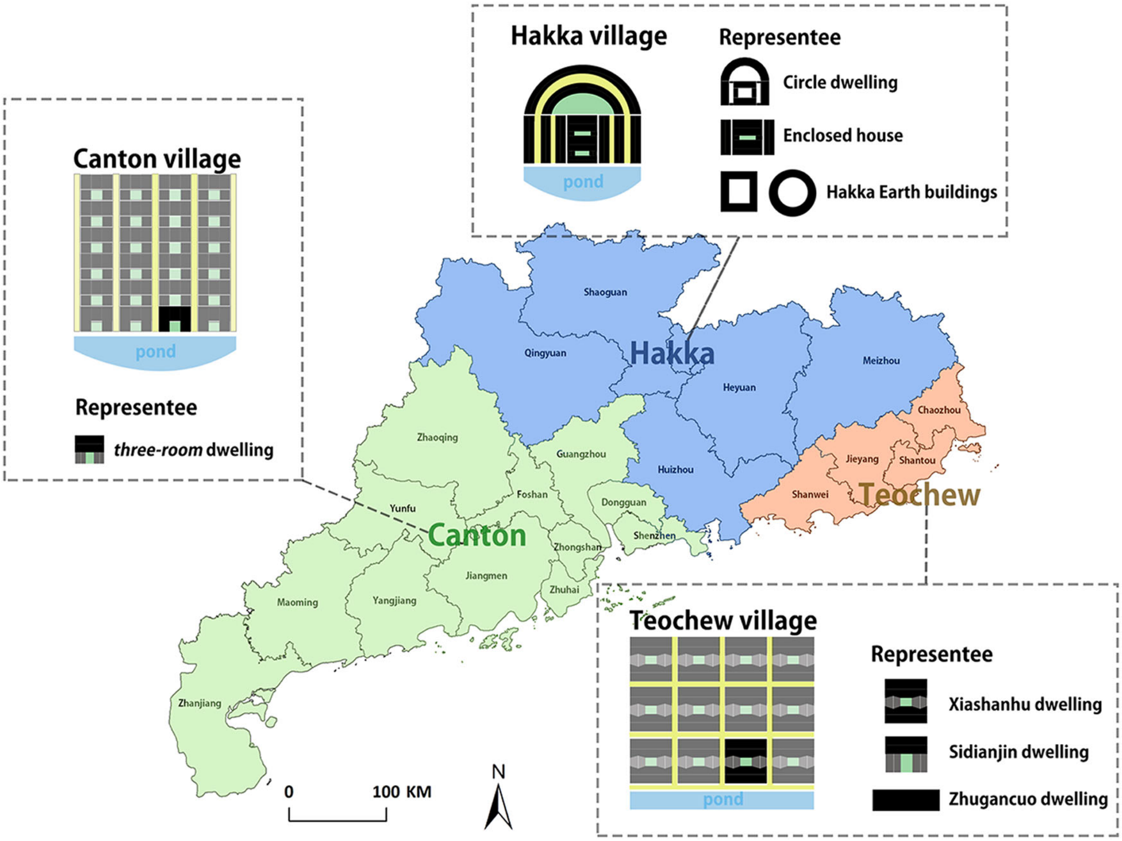


Fig. 1 Approximate distribution of three major ethnic groups in Guangdong, China based on conventional surveys and field studies. Three bounding boxes show their classic footprints of the building layouts and the representees according to domain knowledge.

whole earth, with a resolution of 0.3 m per pixel and three channels (Red, Green, and Blue). The POIs data is used to determine the location of the traditional villages, allowing us to download satellite imagery at a smaller scale.

Methodology. We propose the novel XDR framework that allows infusing explicit domain knowledge to explain the AI model's outputs. The XDR starts with a trained image segmentation model that learns rural building footprints from large amounts of remote-sensing images. More details about the building segmentation model are addressed in the "Pre-condition" section. Afterward, as illustrated in Fig. 2, there are four main steps: (1) *Pyramid Layer Selection* selects the feature maps that are relevant to the XAI problem. Specifically, the proper layer generated by the Feature Pyramid Network (FPN) can better explain the historical building layout in terms of the spatial dimension, (2) *Building- and Village-Scale Feature Extraction* transforms image-scale feature maps to building- and village-scale features, (3) *Infusion of Domain Knowledge* aims to quantitatively estimate the importance of different features on the differentiation of historical village types, and (4) *Proximity Evaluation* clusters different kinds of historical villages in a spatial context and evaluate their proximity relationships. The migration records were also used to validate the proximity relationships and geographical distributions of different kinds of traditional villages. In the second step, we applied a pooling method for aggregating building-scale feature maps to village-scale feature maps. The reason for that is the domain experts labelled villages with different types (see the section "Discussion" for all the village types), rather than labelling on individual buildings. Moreover, historical buildings tend to co-locate in clusters. Therefore, we used village-scale feature maps as the inputs for the third step, allowing the infusion of domain knowledge.

Pre-condition. The trained image segmentation model (called building segmentation model hereafter), which is based on Mask R-CNN (He et al., 2017), is used as the AI black box model to demonstrate the effectiveness of the proposed framework. Mask R-CNN is a well-established model in the field of Convolutional Neural Networks and is widely used in image analysis, e.g., remote sensing image classification. In our previous study, the model is trained to outline building footprints. We pre-trained the model with 1.5 million object instances from the COCO dataset (https://cocodataset.org/). Afterward, it was fine-tuned to recognize building footprints with more than 10,000 annotations. The building footprint training samples cover Guangdong province, China, which are manually collected via visual interpretation from

Table 1 Guangdong's 3 major styles of traditional villages and their approximate location and representees in architectural historiography.

Style of village	Locations	Representees
Canton	Central and west parts of Guangdong	*Three-room dwellings*
Hakka	Mountainous areas of northern Guangdong	*Enclosed houses, Longhouses, Hakka Earth buildings* and *Circle dwellings*
Teochew	East coastline areas	*Xiashanhu dwellings, Sidianjin dwellings*, and *Longhouses*

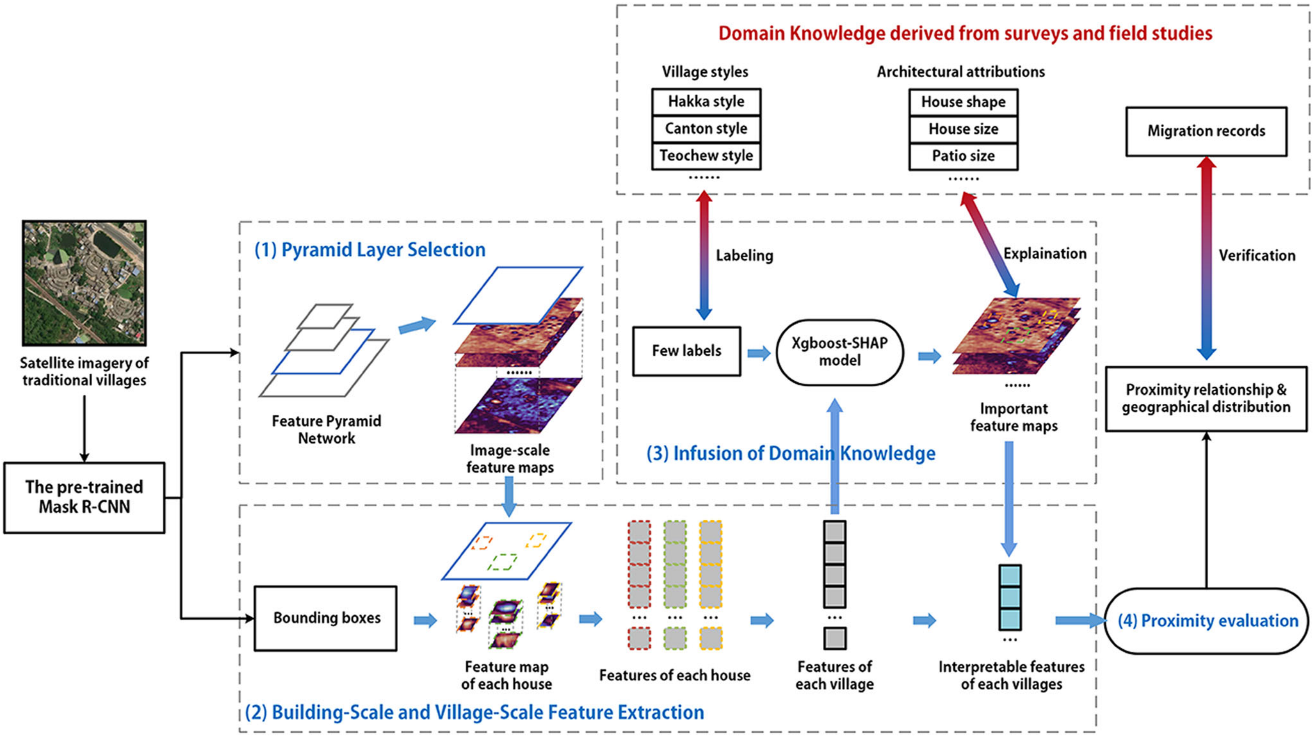

Domain Knowledge derived from surveys and field studies
Village styles
Hakka style
Canton style
Teochew style
......
Architectural attributions
House shape
House size
Patio size
......
Migration records
Satellite imagery of traditional villages
The pre-trained Mask R-CNN
(1) Pyramid Layer Selection
Feature Pyramid Network
......
Image-scale feature maps
Labeling
Explaination
Verification
Few labels
Xgboost-SHAP model
......
Important feature maps
(3) Infusion of Domain Knowledge
Proximity relationship & geographical distribution
Bounding boxes
Feature map of each house
Features of each house
Features of each village
Interpretable features of each villages
(4) Proximity evaluation
(2) Building-Scale and Village-Scale Feature Extraction

Fig. 2 The workflow of the proposed XDR framework. It illustrates how the satellite imagery of a traditional villages is processed to recognize its ethnic style. The XDR framework includes the data-driven part (blue text) and the domain knowledge part (red text). And the red-blue gradient arrows represent the domain knowledge infusion.

remote sensing imageries. The model has been proven effective in segmenting rural buildings (Li et al., 2021, 2022). This Mask R-CNN can generate three types of outputs, i.e., building instances (building footprints), bounding boxes and the corresponding classes. In this study, we used bounding boxes.

Step 1: Feature pyramid networks layer selection. The building segmentation model uses ResNet and Feature Pyramid Networks (FPN) as the main structure. The ResNet can provide deep convolutional feature maps with rich semantic information to understand objects in imagery. And the FPN can increase the resolution for those rich semantic feature maps to help detect small objects, e.g., buildings in satellite imagery. 5 FPN layers with diverse widths and heights, i.e., {$P2$ (512×512), $P3$ (256×256), $P4$ (128×128), $P5$ (64×64)}, are generated from the ResNet's deep convolutional feature maps. Lin et al. addressed the following equation to help select the proper FPN layer according to the general size of the target object (Lin et al., 2017).

$$k = \left\lfloor k_0 + \log_2\left(\sqrt{wh}/224\right)\right\rfloor \tag{1}$$

Amongst, w and h represent the target object's width and height, respectively. The variable k represents the layer number of FPN for the given target object dimension, while k_0 represents the layer of FPN that could address the target object of dimension at 224*224. Lin et al. recommend $k_0 = 4$. In this study, the maximum width and height of a building can be 90 m, that is, 90,000 pixels in the collected satellite imageries (Lin et al., 2017). Hence, according to Eq. (1), we selected the third layer (P3) of FPN for down-streaming processes (see Fig. 2 for the overall workflow).

Step 2: Building-scale and village-scale feature extraction. The feature maps of the FPN cover the whole satellite image, so other land use features such as waterbodies, vegetation and roads were also included. As a result, the mixed land uses represented in the feature map complicate the dimension reduction process. As domain experts judge the ethnic group of the historical villages based on the traditional buildings, we use building-scale feature maps in the dimension reduction process. To be more specific, we cropped the image-scale 256-channel feature maps (from P3 of the corresponding FPN) according to the buildings' bounding boxes on the image (as shown in Step 2 of Fig. 2). We also aggregated features of all buildings X_{house} into one set of feature layers using Global Average Pooling, represented by X_{village} (see Eqs. (2) and (3)). Amongst, $M_{i,j}$ depicts the feature maps of the i_{th} rowj_{th} column pixel, and $X_{\text{house},k}$ depicts the feature maps of the k_{th} house on the image. m, n represents the width and height (in pixels) of a specific house. p represents the total number of houses in the village.

$$X_{\text{house}} = \frac{\sum_{i=0,j=0}^{m,n} M_{i,j}}{m \times n} \tag{2}$$

$$X_{\text{village}} = \frac{\sum_{k=1}^{p} X_{\text{house},k}}{p} \tag{3}$$

Step 3: Infusion of domain knowledge. The prior domain knowledge impacts the XDR framework in three aspects, i.e., feature importance computation, feature semantic inference, and ethnic proximity assessment.

Feature importance computation. Firstly, we asked domain experts to label the satellite imageries with specific historical village types based on the building style and layout. Afterward, the feature maps of different villages (X_{village}) and the associated village types were used to train the XGBoost (eXtreme Gradient Boosting algorithm)-SHAP (Shapley Additive Explanations) model. The XGBoost-SHAP model comprises two sequential processes. Firstly, we used the XGBoost part to build a tabular data-based model for predicting the types of historical villages. XGBoost is a scalable, distributed gradient-boosted decision tree algorithm. It trains on a dataset D_{train} with n samples (see Eq. (4)), with x_i and y_i represent the feature vector and target class of the i_{th} sample respectively. Equation (5) defines the outputs of XGBoost. The variables K, $\hat{y}_i$, f_k represent the number of decision trees, the prediction, and the function of the k_{th} decision tree, respectively. F stands for the set of all possible CARTs (a set of classification and regression trees). In this study, x_i is the feature vector of the i_{th} village, while y_i is the historical village type. We split the dataset into training (70%) and test set (30%). The resulting model achieves an accuracy of 97% on the hold-out test set.

$$D_{\text{train}} = \left\{\left(x_1, y_1\right), \ldots, \left(x_i, y_i\right), \ldots, \left(x_n, y_n\right)\right\} \tag{4}$$

$$\hat{y}_i = \sum_{k=1}^{K} f_k\left(x_i\right), f_k \in F \tag{5}$$

Second, we used the SHAP part to estimate the importance of different features in the differentiation of village types. SHAP originates from the Shapley value idea in cooperative game theory, estimating the importance of individual inputs based on the weighted aggregation of each local marginal contribution. Equation (6) addresses how to compute the SHAP values φ_i for the i_{th} feature.

$$\varphi_i = \sum_{S \subseteq P,\{i\}} \frac{|S|!(|P| - |S| - 1)!}{|P|!}\left[f_P\left(x_P\right) - f_S\left(x_S\right)\right] \tag{6}$$

Amongst, P represents the number of all features, $S \subseteq P$,{i} depicts the subset of P after removing i_{th} feature. The multiplier f_P (x_P)−f_S (x_S) represent the prediction differences between two XGBoost models; f_P (x_P) represents the outputs of the model trained on all features P, while f_S (x_S) represents the outputs of model that is trained on feature set S. As a result, we derived

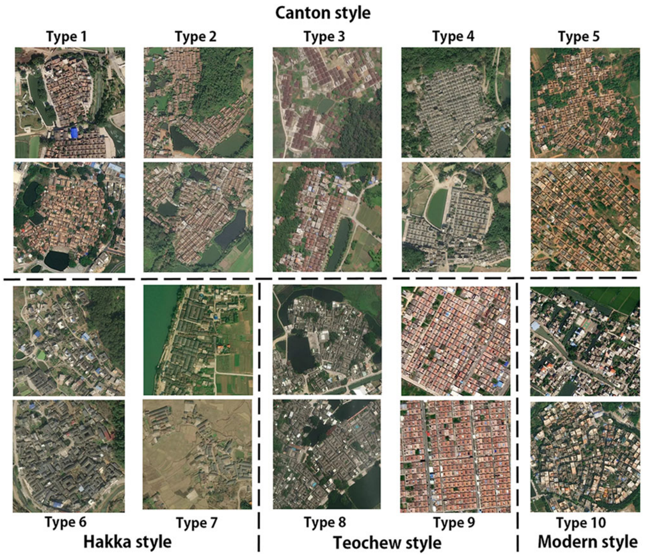


Fig. 3 The example images per village style. The images of Type 1–Type 5 show examples of Canton villages; the images of Type 6 and Type 7 for Hakka villages; the images of Type 8 and Type 9 for Teochew villages; and the images of Type 10 show modern villages.

the importance of different features on the determination of village types.

Feature semantic inference. We visualized the prominent features per village type. Domain experts inferred the architectural semantics of each one of them and explored why those prominent features contribute to the village types. Using these features as predictors, we can label all historical villages with specific styles in Guangdong, inferring their dominant ethnic inheritance. (3) Finally, we compared the proximity and geographical distributions of the latent ethnic inheritance against the migration records (derived from surveys and field studies) to verify the result. This process also enhances explicit domain knowledge, allowing domain experts to identify and locate the neglected historical villages.

Ethnic proximity assessment. More specifically, the infusion process starts with curated examples of nine distinctive village styles from three ethnic groups (see Fig. 3). Moreover, as modern buildings have mixed with historical buildings, the domain experts also concretized one additional style for the modern buildings. Based on the annotation criteria, the domain experts labelled seven to nine images for each style, accumulating 84 example images. The data is used to train and validate the XGBoost-SHAP model. For the robustness of the XGBoost-SHAP model, we applied a random sampling strategy. To be more specific, we randomly selected 60% of the samples per historical village type as the inputs of the XGBoost-SHAP model, deriving the mean SHAP value per input feature. Such a process was repeated 500 times. The mean of the average SHAP values per feature is regarded as importance in the determination of village types X_{village}.

Once we identified the prominent features $M_{\text{semantics}}$ through the XGBoost-SHAP model, we aggregated them based on the building-scale features and derived the village-scale feature vectors ($X_{\text{semantics}}$). This dimension reduction process allows us to extract the most prominent *n* features from the 256-channel P3 feature maps.

Step 4: Proximity evaluation. This step aims to obtain the proximity relationships and geographical distributions of all 10 types of villages based on the feature maps $X_{\text{semantics}}$. It starts with computing the *cosine similarity* between any two villages

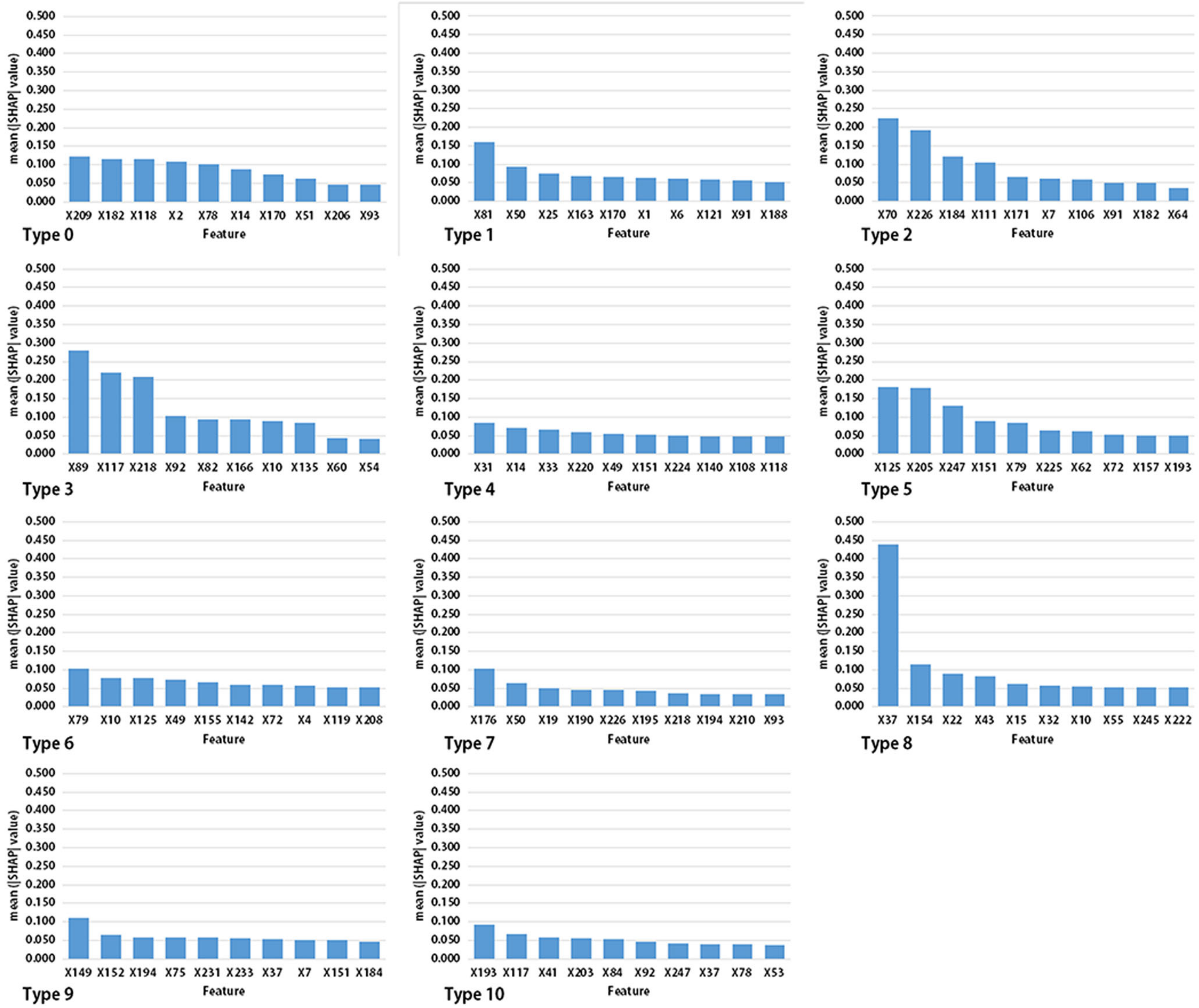


Fig. 4 The importance ranking of semantic features on each building type via their SHAP scores. The higher the SHAP score of a feature, the more important it is for distinguishing that type. Hence, the features ranked first for the ethnic types are the most useful.

(see Eq. (7)).

$$\cos\theta = \frac{X_m \cdot X_n}{|X_m| \cdot |X_n|} = \frac{\sum_{i=1}^{11} x_{m,i} \times x_{n,i}}{\sqrt{\sum_{i=1}^{11} x_{m,i}^2} \times \sqrt{\sum_{i=1}^{11} x_{n,i}^2}} \tag{7}$$

Amongst, X_m and X_n represent the feature maps $X_{\text{semantics}}$ of village m and n, respectively. $x_{m,i}$ depicts the $i \in (1\ldots11)$ feature vector of village m. Afterward, we built a graph of villages based on the similarity matrix. Moreover, we used Gephi, a graph visualization software, to visualize the graph. K-means clustering method is used to cluster the villages based on the similarity matrix.

Results

The prominent features determining the village types. We found the 11 most prominent features via the XGBoost-SHAP method from deep convolutional features that can impact the decision of village types (Fig. 4). M_{209} stands for the 209th feature map of the 256-channel feature maps generated by FPN of Mask R-CNN.

$$M_{\text{semantics}} = \left\{ \begin{matrix} M_{209}, M_{81}, M_{70}, M_{89}, M_{31}, M_{125}, M_{79}, \\ M_{176}, M_{37}, M_{149}, M_{193} \end{matrix} \right\} \tag{8}$$

By aggregating these features based on the building scale and then the village scale, we derived the feature maps per village as below. The corresponding SHAP values of each feature per building type are shown in Fig. 3. The distributions of the prominent features' $M_{\text{semantics}}$ vary significantly across all building types. For example, X_{37} contributes the most impact in building Type 8 (i.e., one of the two Teochew styles), but marginally in Type 10 (i.e., the modern style). We explain the semantics meaning of each prominent feature in the next section through domain knowledge. In addition, the $X_{\text{semantics}}$ of all villages in this study can be found in the supplementary information file - Supplementary Dataset S1.

$$X_{\text{semantics}} = \left\{ \begin{matrix} X_{209}, X_{81}, X_{70}, X_{89}, X_{31}, X_{125}, X_{79}, \\ X_{176}, X_{37}, X_{149}, X_{193} \end{matrix} \right\} \tag{9}$$

The architectural semantics of different feature maps. Each feature of the prominent feature set $M_{\text{semantics}}$ should reflect one or more architectural characteristics. We visualized each feature map M_i by overlaying the corresponding satellite images and asking the domain experts to infer the semantic meaning. As shown in Fig. 5, we overlayed the activation map of M_{89} with satellite images of different village types. From the fourth column, we can see activations are high in the patio part of a building, and low in the surrounding parts. As a result, domain experts inferred M_{89} is sensitive to the patio. Patio is an open space inside a house, linking different functional areas, providing natural daylight, and harvesting stormwater. However, it is rare in modern buildings but popular in historical buildings. Therefore, domain experts believe they can use it to determine the locations of all historical buildings in the case study area.

The other features reflect the other specific architectural characteristics, e.g., the size (see Supplementary Fig. S1), the length (see Supplementary Fig. S2) and the direction (see Supplementary Fig. S3) of buildings. The features for building direction are particularly interesting, as activations of those features are asymmetric (see Supplementary Fig. S3 for more details). However, these features allow us to identify Hakka buildings, which always form in curly structures, i.e., *Circle dwellings*.

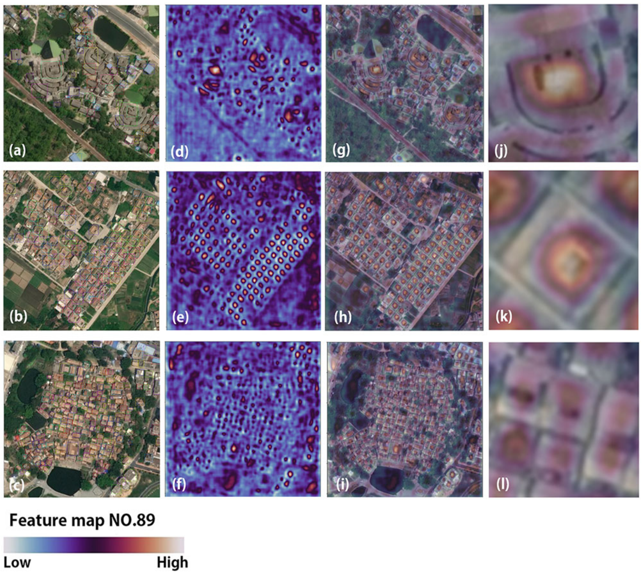


Fig. 5 The activation of the feature map is sensitive to the patio of a building. Panels **a–c** show satellite images with different village styles. The building bounding boxes were outlined. Panels **d–f** show the corresponding activation map of the images in the same row. Panels **g–i** overlays the images with the corresponding activation map. Panels **j–l** enlarge images of Panels **g–i** to show the activations in detail.

The mixed villages and spatial distributions. The prominent features derived from the XDR framework allow us to determine the styles of villages at scale. To be more specific, we computed the similarity between any pair of villages and grouped them into eight clusters (see Section 5.4), building a village network as shown in Fig. 6. Afterwards, we let the domain experts annotate the styles based on the corresponding satellite imageries. The styles are defined as *Hakka village (Shaoguan–Qingyuan type)*, *Canton–Hakka mixed village*, *Canton village*, *Canton–Teochew mixed village*, *Hakka–Teochew mixed village, Hakka village (Meizhou type)*, *Teochew village*, and *Modern village*, according to their classic examples.

The data-driven clustering result interestingly illustrates how the three ethnic groups impact the building styles of each other. Between any two styles that are dominated by one single ethnic group, we can see the mixed style emerging. For example, between the *Hakka villages (Shaoguan–Qingyuan type)* and the *Canton villages*, we observed the mixed-style *Canton–Hakka mixed villages*. The middle of the network is the *Modern village*, which might be a product of urbanization. To validate this assumption, we mapped all these villages on the case study area (as shown in Fig. 7). The villages of the same ethnic style are co-located geographically. For example, the *Canton villages*, *Hakka villages (Meizhou type),* and *Teochew villages* are dominant in the central, eastern, and eastern coastal areas of Guangdong. On the other hand, *mixed villages* distribute across the whole area. Here is the summary of the village distribution.

- *Canton* and *Hakka villages* are the most dominant historical village styles. *Canton villages* distribute broadly in the western part of Guangdong, including Guangzhou, Dongguan, Zhongshan, Zhaoqing, Jiangmen and Yangjiang. *Hakka villages (Meizhou type)* are dominant in the eastern part of Guangdong, including the Meizhou and Heyuan. The *Hakka villages (Shaoguan–Qingyuan type)* are located in the Shaoguan and Qingyuan cities. The *Canton–Hakka* villages are distributed broadly across Guangdong, surrounding the *Canton villages* in the middle.

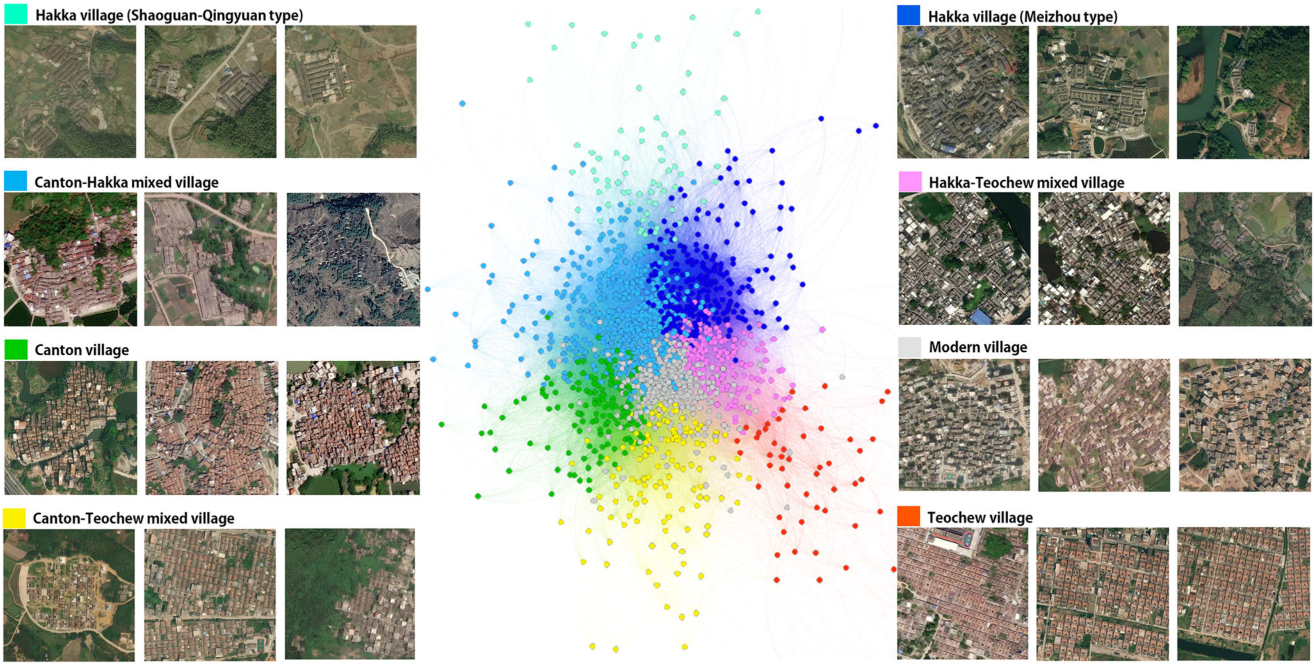


Fig. 6 The proximity network and the classical examples of each village style. Dots of the same color are grouped together to form clusters. This indicates that the distinguishing features of different ethnic types are captured. These classical examples of each styles are given by the data-driven clustering algorithm. And they all fit to their corresponding ethnic styles according to the domain knowledge.

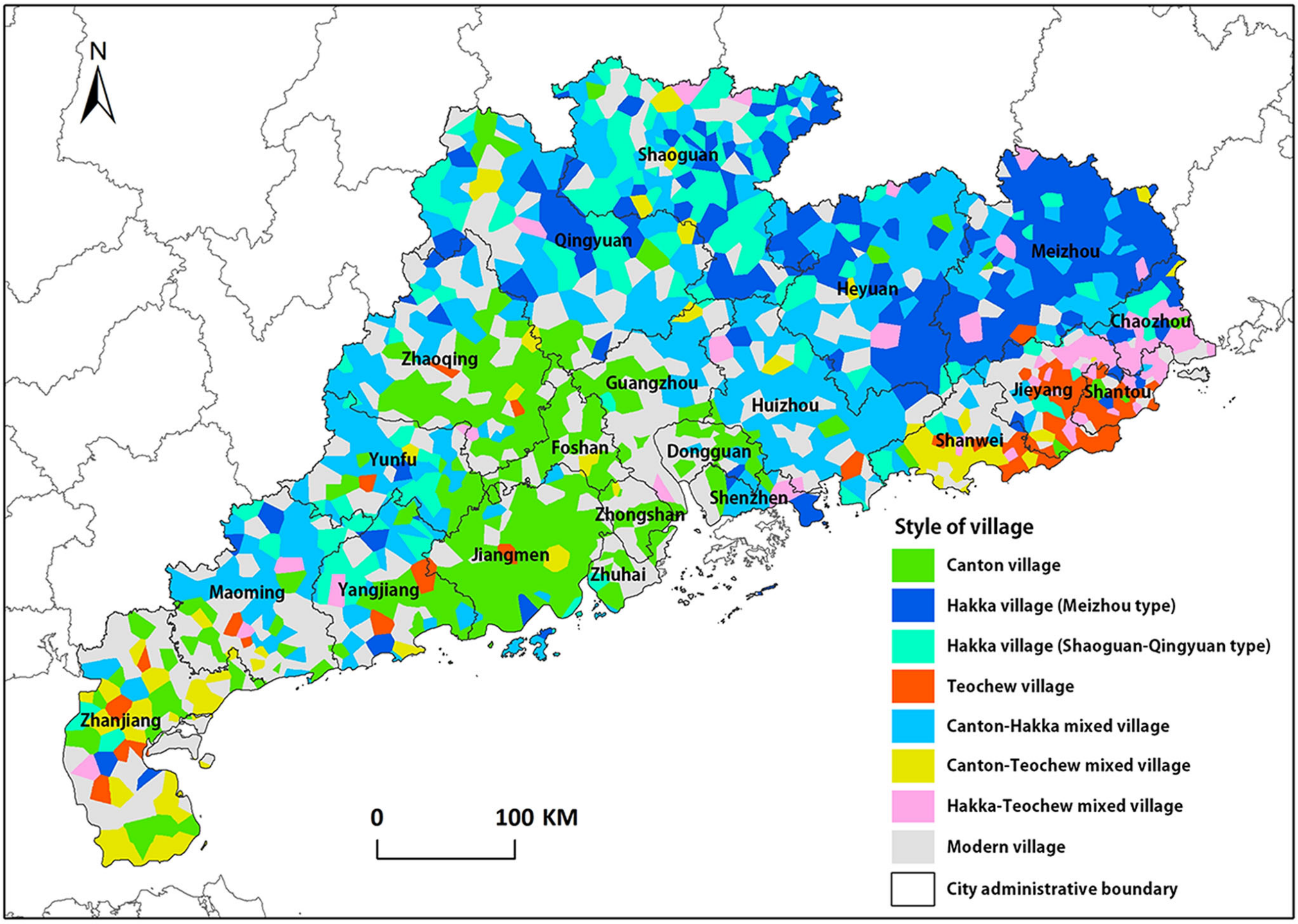
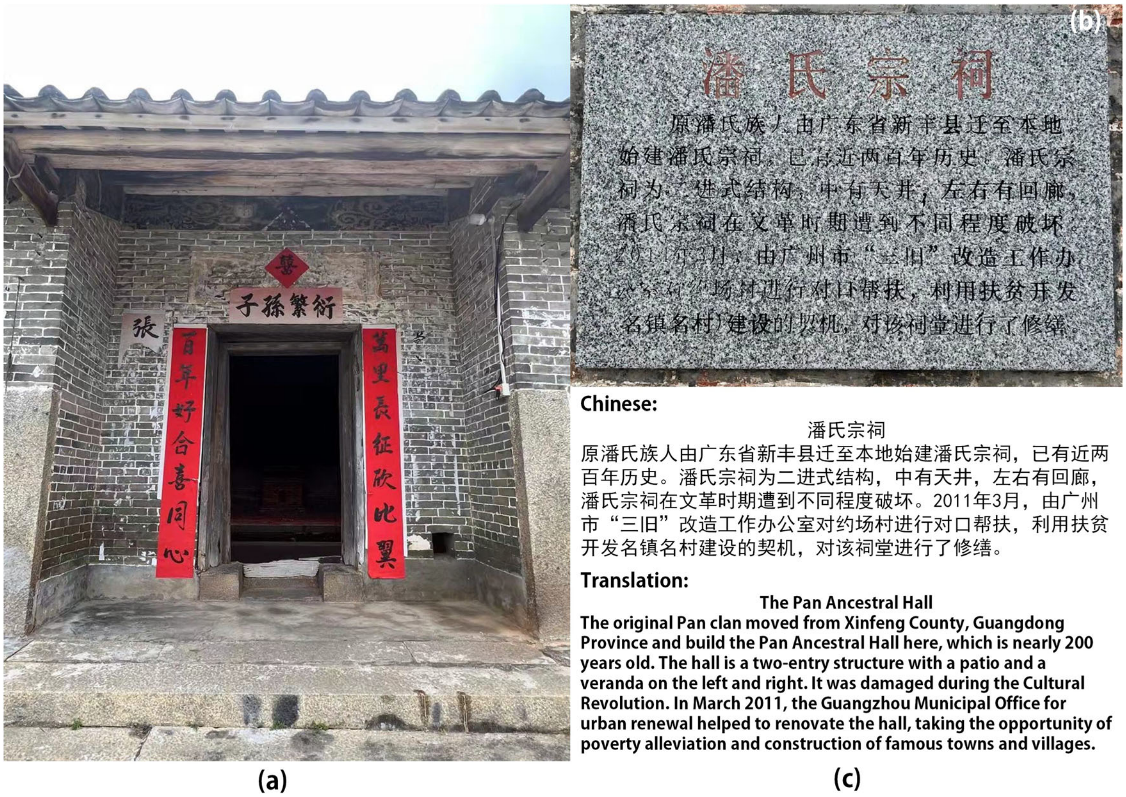


Fig. 7 The Voronoi diagram of diverse village styles. Voronoi diagram is used to better visualize the geographical pattern of the collected samples. It reveals the geographical agglomeration of the ethnic styles of village dwellings.



Chinese:

潘氏宗祠

原潘氏族人由广东省新丰县迁至本地始建潘氏宗祠，已有近两百年历史。潘氏宗祠为二进式结构，中有天井，左右有回廊，潘氏宗祠在文革时期遭到不同程度破坏。2011年3月，由广州市“三旧”改造工作办公室对约场村进行对口帮扶，利用扶贫开发名镇名村建设的契机，对该祠堂进行了修缮。

Translation:

The Pan Ancestral Hall

The original Pan clan moved from Xinfeng County, Guangdong Province and build the Pan Ancestral Hall here, which is nearly 200 years old. The hall is a two-entry structure with a patio and a veranda on the left and right. It was damaged during the Cultural Revolution. In March 2011, the Guangzhou Municipal Office for urban renewal helped to renovate the hall, taking the opportunity of poverty alleviation and construction of famous towns and villages.

Fig. 8 The Pan Ancestral Hall in Yuechang Village of Zengcheng County, Guangzhou. a The photo of the main entrance of the ancestral hall. **b** The stele in the ancestral hall recording the migration record of the village. **c** The text on the stele and its translation.

- In the eastern part of Guangdong, *Teochew villages* and *the related-mixed villages* are popular. The Teochew villages are located in Shantou and Jieyang, while the Hakka- Teochew villages are located in Chaozhou and partial Jieyang (the pink area in the eastern of Guangdong). Teochew–Canton villages are located in Shanwei, the coastal areas in the east, surrounded by the Teochew and Canton villages.
- The western part of Guangdong comprises villages of diverse styles. Amongst, Canton villages and Hakka villages *(Shaoguan–Qingyuan type)* are located in Yunfu, Yangjiang and Maoming. While in Zhanjiang, Canton villages, Teochew villages, Hakka villages are highly blended. Canton–Hakka mixed villages are frequently seen in the far West Guangdong.

Investigating a migration case. The proposed XDR framework generates a series of interesting knowledge that is new to the domain experts, such as the distribution map (see Fig. 7), offering a low-cost survey with an extensive scale for human geography research. Domain experts found one *Canton–Hakka village* area in Fig. 7 interesting. The area is located in the eastern part of Guangzhou, which is supposed to be dominated by Canton villages and Mixed villages. Domain experts carried out a field study on that *Canton–Hakka village* area and found an important piece of recorded history in the Yuechang Village of Zengcheng County, Guangzhou (see Fig. 8). The stele outside of the Pan Ancestral Hall states that the Pan clan migrated from Xinxing County, Shaoguan to Zengcheng County, Guangzhou about 200 years ago. Xinxing County is dominated by Hakka villages.

Afterward, we investigated the migration history of Zengcheng County based on the distribution map and the corresponding satellite images. Xinfeng County (as mentioned in the stele) is located in the southern part of Shaoguan City, which is 60–100 km away from Zengcheng County of Guangzhou City. According to the historical records, this migration of the Pan clan took place during the fourth period of Hakka migration due to wars and population explosion in the early Qing Dynasty. This migration route from Xinfeng County to Zengcheng County is one of the main routes of the fourth migration (Cohen, 1968; Leong et al., 1997; Lowe, 2012). As shown in Fig. 9, Yuechang Village and Dongdong Village show similar Hakka building styles as Xinfeng County, which is different from the dominant Canton style in the local villages. For example, *Enclosed houses* and *Longhouses* can be found in that area. We assume the style of the historical buildings in these two villages could be impacted by both the styles of the clan's origin area and the local area.

Ablation study of proposed explainable dimensionality reduction. To validate the performance of the proposed XDR framework, we compared the village networks between different methodologies. That includes the computation based on *setting* (1) the original 256-channel P3 feature maps, (2) the 11 principal components from PCA analysis of the 256-channel P3 feature maps, (3) the 256-channel P3 feature maps at the village scale, (4) the 11 principal components from PCA analysis of the 256-channel P3 feature maps at the village-scale, and (5) the feature maps from the XDR framework. To be more specific, in *setting* (1), the original feature maps of a given image are averaged at the image level, which is depicted as the feature vector of the given image. In *setting* (2), the image-scale feature vectors are converted into the 11 principal components via PCA analysis. In *setting* (3), the feature maps of each building bounding boxes at the same village are averaged, and these feature maps are village-scale feature vectors. In *setting* (4) all village-scale feature vectors were converted to 11 principal components based on the PCA analysis. *Setting* (5) presents the village-scale feature vectors using the XDR framework.

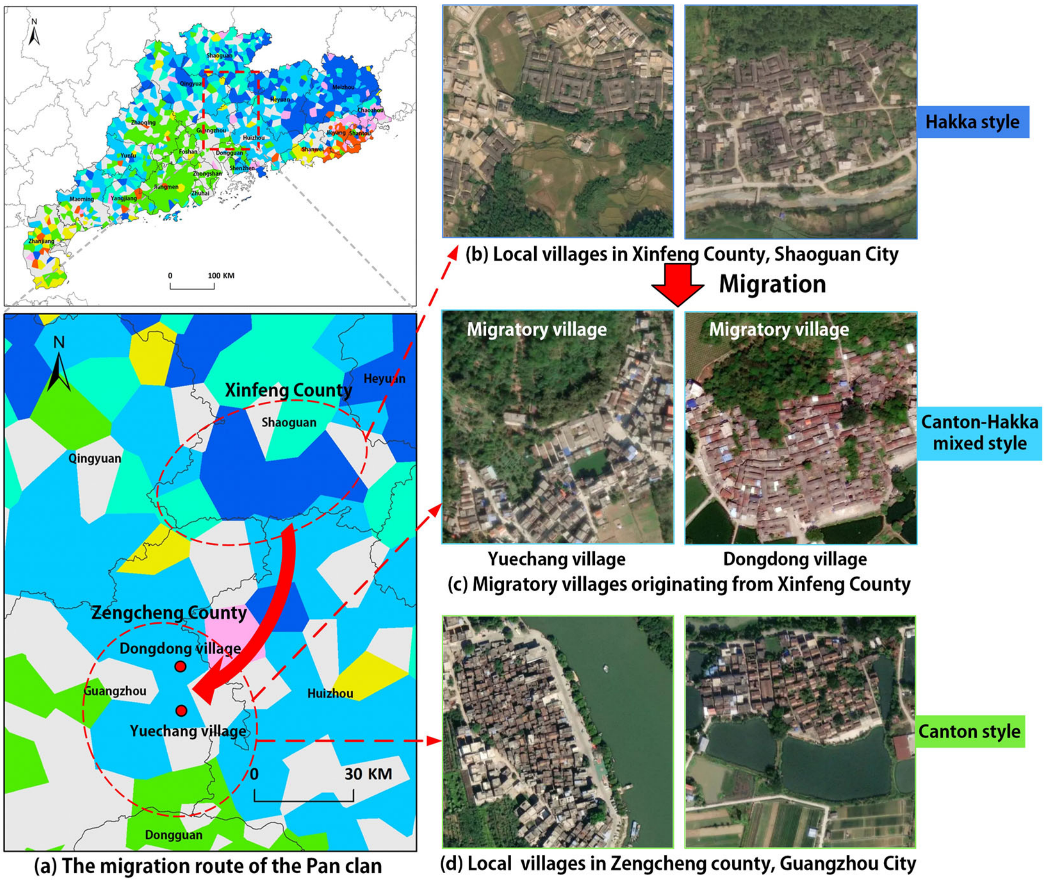


Fig. 9 The migration route of the Pan clan and the village layouts shown in the satellite images. a The migration route of the Pan clan (the red arrow) is drawn based on the historical records of the Pan Ancestral Hall. **b** The local villages at the start of the migration route (Xinfeng County, Shaoguan City) show Hakka style in the satellite imagery. **d** The local villages at the end of the migration route (Zengcheng County, Guangzhou City)are Canton style in the satellite imagery. And (**c**) the migratory villages show Canton-Hakka mixed style in the satellite imagery.

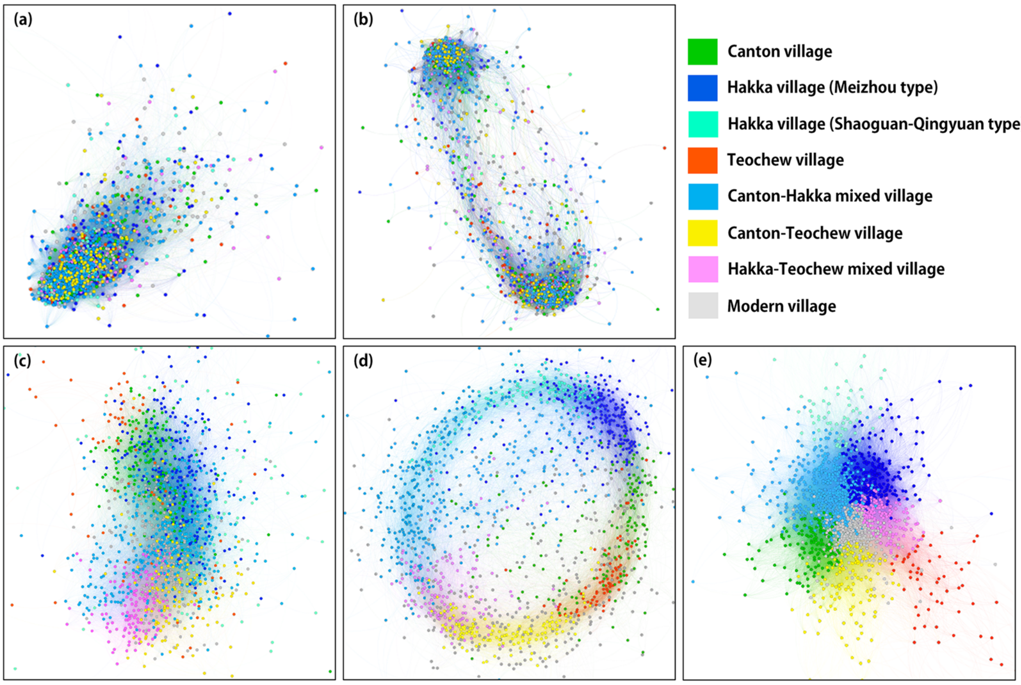


Fig. 10 The village networks were computed using different methods. The node colour represents the cluster group derived from panel (**e**). Panel **a** shows the village network computed by the original 256-channel feature maps. Panel **b** shows the result based on the 11 principal components from PCA analysis of the 256-channel feature maps. Panel (**c**) shows the result using the 256-channel feature maps at the building scale. Pane **d** is the 11 principal components from PCA analysis of the 256-channel feature maps at the building scale. Panel **e** is the result using the XDR framework.

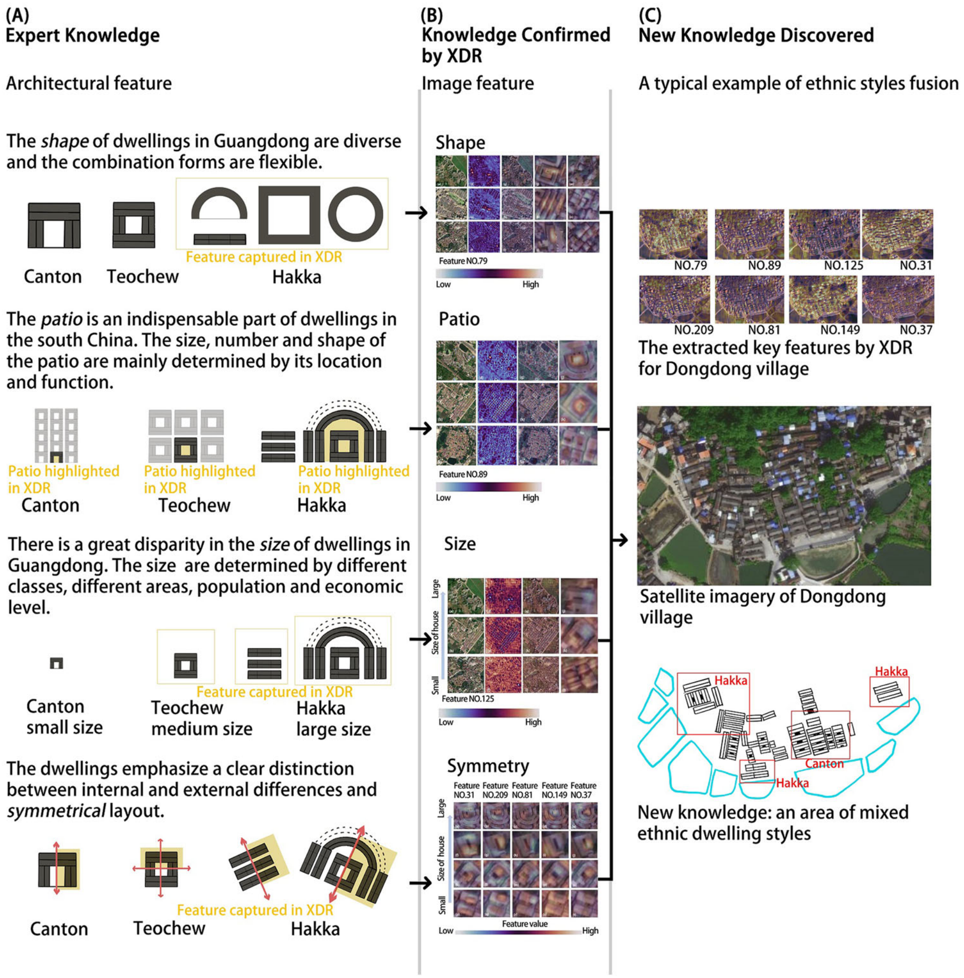

(A)
Expert Knowledge
Architectural feature
The shape of dwellings in Guangdong are diverse and the combination forms are flexible.
Feature captured in XDR
Canton
Teochew
Hakka
The patio is an indispensable part of dwellings in the south China. The size, number and shape of the patio are mainly determined by its location and function.
Patio highlighted in XDR
Canton
Patio highlighted in XDR
Teochew
Patio highlighted in XDR
Hakka
There is a great disparity in the size of dwellings in Guangdong. The size are determined by different classes, different areas, population and economic level.
Feature captured in XDR
Canton small size
Teochew medium size
Hakka large size
The dwellings emphasize a clear distinction between internal and external differences and symmetrical layout.
Feature captured in XDR
Canton
Teochew
Hakka
(B)
Knowledge Confirmed by XDR
Image feature
Shape
Feature NO.79
Low
High
Patio
Feature NO.89
Low
High
Size
Large
Size of house
Small
Feature NO.125
Low
High
Symmetry
Feature NO.31
Feature NO.209
Feature NO.81
Feature NO.149
Feature NO.37
Large
Size of house
Small
Feature value
Low
High
(C)
New Knowledge Discovered
A typical example of ethnic styles fusion
NO.79
NO.89
NO.125
NO.31
NO.209
NO.81
NO.149
NO.37
The extracted key features by XDR for Dongdong village
Satellite imagery of Dongdong village
Hakka
Hakka
Canton
Hakka
New knowledge: an area of mixed ethnic dwelling styles

Fig. 11 How XDR confirms expert knowledge and helps discover new knowledge. Panel **a** addresses what expert knowledge is used for infusion. The expert knowledge that is confirmed by the XDR framework is highlighted in yellow. Panel **b** addresses which expert knowledge is confirmed by the XDR. The confirmed architectural feature is highlighted in different feature maps. Panel **c** addresses how the XDR helps discover new knowledge by presenting an example of the discovery of a village of mixed ethnic styles.

As shown in Fig. 10, the village network based on the 256-channel P3 feature maps doesn't reveal a clear structure (panel a). The PCA-based village network shows two clusters (panel b). However, we do not know the semantics meaning of the principal components as they are aggregations between all feature maps. For the third methodology, we can see a clustered structure (panel c), but it's not as clear as the result of the XDR framework. Panel d shows a ring-shaped structure with clear clusters. That means PCA combined with the building-scale feature maps is also a good alternative. However, we found that Canton villages are located between the Teochew and Hakka villages, which is different from the existing findings as shown in Fig. 7. Moreover, we do not know the semantics meaning of each principal component as the other PCA-based methodologies. Another finding is the methodologies using the building-scale feature maps are more robust than the other methodologies (XDR framework is also based on building-scale feature maps). The reason for that is the other land uses information that could contaminate the analysis is excluded.

Discussion

This study proposed the XAI method specific for dimension reduction and domain knowledge infusion, allowing us to enhance our domain knowledge from the machine's tacit knowledge. The benefit of domain knowledge development is twofold: the confirmation of the domain knowledge and the discovery of new knowledge. In Fig. 11, we present how XDR confirms expert knowledge and helps discover new knowledge. The domain experts brought forward four main aspects of architectural knowledge (Fig. 11a). Some of them were confirmed by the XDR. To be more specific, the *shape* of the Hakka's dwellings, all four kinds of the *patio*, the *size* of the Teochew and Hakka's dwellings, and all *symmetrical* layouts were confirmed at different feature maps in the XDR (Fig. 11b). Based on the feature maps that are relevant to the expert knowledge, we found some villages where the dwellings have mixed ethnic styles. This kind of mixed-ethnic historic village is generally undocumented. The discovery benefits domain experts in terms of the understanding of cultural integration and human migration. The proposed XDR framework makes contributions in three other aspects as below.

(1) The proposed framework benefits from the SHAP value and attention mechanism. Firstly, we used the SHAP value to measure the importance of different features and summarized their positive and negative impacts on the determination of village types. Visualization of the SHAP value allows us to derive the semantics meaning. Moreover, the proposed XDR framework builds upon a Mask R-CNN-based building detection model. The model itself attends to the building details, allowing the down-streaming tasks in the XDR framework to focus on buildings and be isolated

from other noise, e.g., features related to the other land uses information. Also, the prior knowledge is particularly focused on architectural characterization, so the building-related model features can be aligned perfectly.

(2) The infusion of domain knowledge process in the proposed XDR framework does not require a large number of human labels. That is significant to human geography research, as there is a lack of publicly available training samples. The framework is of few-shot learning capability.

(3) Most importantly, the proposed XDR framework can be applied in village classification and spatial proximity analysis at a large scale thanks to the increasing availability of high-quality satellite imagery, offering a new perspective for human geography research. Humans have a profound history of migration for better livelihoods (De Haan, 1999). The migration encourages cultural integration, reflected in the mixed building styles of different ethnic cultures (Burmeister, 2000). The proposed XDR framework allows us to understand cultural integration by mapping building styles at scale. Furthermore, we could discover historic villages and even undocumented migration routes. The XDR framework is also applicable in other countries.

There are limits to the XDR framework. Firstly, the framework relies on the accuracy of the bounding box of the Mask R-CNN model. If the model performs poorly in some areas, the performance of the XDR framework will be compromised. Second, there is a risk of losing secondary information via the current XDR framework. Compared to the conventional method of dimensionality reduction, the proposed XDR considers the most important features for each category and thus can retain the primary crucial information. However, the second or third most important features for each category may also affect the results. And the current experiments have not considered those features in order to ensure the conciseness of the domain knowledge part of the content. Integrating the top three important features can be considered in future studies.

Data availability

The raw data of satellite imageries are available in MapQuest (www.mapquest.com), a satellite imagery provider in the United States. The villages' features generated by the XDR framework and analysed in the "Results" section are available in the supplementary information file, Supplementary Dataset S1, online.

Received: 23 June 2022; Accepted: 4 January 2023;
Published online: 26 January 2023

References

Angwin J, Larson J, Mattu S, Kirchner L (2022) Machine Bias*. Ethics of Data and Analytics 254–264. https://doi.org/10.1201/9781003278290-37

Brundage M, Avin S, Wang J et al. (2020) Toward trustworthy AI development: mechanisms for supporting verifiable claims. Preprint at https://doi.org/10.48550/arxiv.2004.07213

Burmeister S (2000) Archaeology and migration: approaches to an archaeological proof of migration. Curr Anthropol 41(4):539–567

Cohen ML (1968) The Hakka or "Guest People": dialect as a sociocultural variable in Southeastern China. Ethnohistory 15:237–292

Comon P (1994) Independent component analysis, a new concept? Signal Process 36:287–314. https://doi.org/10.1016/0165-1684(94)90029-9

De Haan A (1999) Livelihoods and poverty: the role of migration—a critical review of the migration literature. J Dev Stud 36(2):1–47

Devlin J, Chang M-W, Lee K et al. (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805

Díaz M, Johnson I, Lazar A et al. (2018) Addressing age-related bias in sentiment analysis. In: Proceedings of the 2018 CHI conference on human factors in computing systems, Association for Computing Machinery, New York, 21–26 April 2018

Dombrowski AK, Alber M, Anders CJ et al. (2019) Explanations can be manipulated and geometry is to blame. In: Advances in neural information processing systems, The MIT Press, 30 November 2019

Fuentes JM, García AI, Ayuga E, Ayuga F (2011) The development of the flour-milling industry in Spain: analysis of its historical evolution and architectural legacy. J Hist Geogr 37:232–241. https://doi.org/10.1016/J.JHG.2010.10.002

Gao J, Wu B (2017) Revitalizing traditional villages through rural tourism: a case study of Yuanjia Village, Shaanxi Province, China. Tour Manag 63:223–233. https://doi.org/10.1016/J.TOURMAN.2017.04.003

Grigorescu S, Trasnea B, TC-J of F, 2020 undefined (2020) A survey of deep learning techniques for autonomous driving. Wiley Online Libr 37:362–386. https://doi.org/10.1002/rob.21918

He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.

Islam SR, Eberle W, Ghafoor SK (2020a) Towards quantification of explainability in explainable artificial intelligence methods. In: Roman Barták, Eric Bell (ed). Proceedings of the 33rd International Florida Artificial Intelligence Research Society Conference, Florida, 17–20 May 2020

Islam SR, Eberle W, Ghafoor SK et al. (2020b) Domain knowledge aided explainable artificial intelligence for intrusion detection and response. In Martin A, Hinkelmann K et al. (Eds.): Proceedings of the AAAI 2020 Spring Symposium on Combining Machine Learning and Knowledge Engineering in Practice (AAAI-MAKE 2020). Stanford University, Palo Alto, 23-25 March 2020

Kambhampati S (2021) Polanyi's revenge and AI's new romance with tacit knowledge. Commun ACM 64:31–32. https://doi.org/10.1145/3446369

Lapuschkin S, Wäldchen S, Binder A et al. (2019) Unmasking Clever Hans predictors and assessing what machines really learn. Nat Commun 10:1–8. https://doi.org/10.1038/s41467-019-08987-4

Leong S-T, Wright T, Skinner GW (1997) Migration and ethnicity in Chinese history: Hakkas, Pengmin, and their neighbors. Stanford University Press

Li X, Xu W, Huang Y et al. (2022) Spatial distribution of rural building in China: remote sensing interpretation and density analysis. Acta Geogr Sin 77:835–851

Li Y, Xu W, Chen H et al.(2021) A novel framework based on mask R-CNN and histogram thresholding for scalable segmentation of new and old rural buildings. Remote Sens 13:1070. https://doi.org/10.3390/RS13061070

Lin T-Y, Dollar P, Girshick R et al. (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), IEEE, San Francisco 21–26, July 2017

Lombrozo T (2006) The structure and function of explanations. Trends Cogn Sci 10:464–470. https://doi.org/10.1016/J.TICS.2006.08.004

Lowe KD (2012) Heaven and earth—sustaining elements in Hakka Tulou. Sustainability 4:2795–2802. https://doi.org/10.3390/su4112795

Lu K, Mardziel P, Wu F et al. (2020) Gender bias in neural natural language processing. Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics) In Logic, Language, and Security 12300 LNCS. pp. 189–202. Springer, Cham.

Lu Q (2008) Local dwellings in Guangdong. China Architecture & Building Press, Beijing

Lu Y (1981) Local dwellings in Guangdong. Archit J 09:29–36

Lu Y (2007) Fifty years of research on Chinese folk house. Archit J 11:67–69

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In Guyon I, Von Luxburg U et al. (eds): Advances in neural information processing systems. Long Beach, 4–9 December 2017

Mehrabi N, Morstatter F, Saxena N et al. (2021) A survey on bias and fairness in machine learning. dl.acm.org 54: https://doi.org/10.1145/3457607

Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38

Polanyi, M. (2009). The tacit dimension. In Knowledge in organizations. Routledge, pp 135–146

Qin RJ, Leung HH (2021) Becoming a Traditional village: heritage protection and livelihood transformation of a Chinese Village. Sustainability 13:2331. https://doi.org/10.3390/SU13042331

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, Association for Computing Machinery San Francisco, 13–17 August 2016

Roscher R, Bohn B, Duarte MF, Garcke J (2020) Explainable machine learning for scientific insights and discoveries. IEEE Access 8:42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199

Ruggiero G, Parlavecchia M, Dal Sasso P (2019) Typological characterisation and territorial distribution of traditional rural buildings in the Apulian territory (Italy). J Cult Herit 39:278–287. https://doi.org/10.1016/J.CULHER.2019.02.012

Russakovsky O, Deng J, Su H et al. (2015) ImageNet large scale visual recognition challenge. Int J Comput Vision 115:211–252. https://doi.org/10.1007/S11263-015-0816-Y

Samek W, Wiegand T, Müller KR (2017) Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296.

Situ S (2001) Lingnan historical and human geography: a comparative study of Guangfu, Hakka and Fulao Ethnic Group. Sun Yat-sen University Press, China, Guangzhou

Wang J, Tuyls J, Wallace E, Singh S (2020) Gradient-based analysis of NLP models is manipulable. findings of the Association for Computational Linguistics Findings of ACL: EMNLP 247–258.Preprint at https://doi.org/10.48550/arxiv.2010.05419

Winkler JK, Fink C, Toberer F et al. (2019) Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol 155:1135–1141. https://doi.org/10.1001/JAMADERMATOL.2019.1735

Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemometr Intell Lab Syst 2:37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Yang G, Ye Q, Xia J (2022) Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. Inf Fusion 77:29–52

Zanfi F, Merlini C, Giavarini V, Manfredini F (2020) A portrait of Italian 'Family houses': diversified heritage in a redefined territorial and demographic context. City Territ Archit 7:1–16. https://doi.org/10.1186/S40410-020-00125-8/FIGURES/12

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-023-01505-4.

Correspondence and requests for materials should be addressed to Xun Li.