



Classifying arbitrary urban morphotypes: a typology-aligned, open-world approach

Dongsheng Chen , Yao Yao , Xun Li , Jinzhou Cao , Yu Gu & Liqiu Meng

To cite this article: Dongsheng Chen , Yao Yao , Xun Li , Jinzhou Cao , Yu Gu & Liqiu Meng (07 Oct 2026): Classifying arbitrary urban morphotypes: a typology-aligned, open-world approach, Annals of GIS, DOI: [10.1080/19475683.2026.2743305](https://doi.org/10.1080/19475683.2026.2743305)

To link to this article: <https://doi.org/10.1080/19475683.2026.2743305>



© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 07 Oct 2026.



[Submit your article to this journal](#)



[View related articles](#)



[View Crossmark data](#)

Classifying arbitrary urban morphotypes: a typology-aligned, open-world approach

Dongsheng Chen^a, Yao Yao^b, Xun Li^c, Jinzhou Cao^d, Yu Gu^c and Liqiu Meng^a

^aChair of Cartography and Visual Analytics, Technical University of Munich, Munich, Germany; ^bUrbanComp Lab, School of Geography and Information Engineering, China University of Geosciences, Wuhan, China; ^cSchool of Geography and Planning, Sun Yat-sen University, Guangzhou, China; ^dSchool of Artificial Intelligence, Shenzhen Technology University, Shenzhen, Guangdong, China

ABSTRACT

Analyzing urban morphology across different geographical contexts can reveal recurring patterns of urbanization and support the critical evaluation of urban design practices. Existing machine-learning approaches generally follow either supervised classification or unsupervised clustering. Supervised methods retain established morphological knowledge, but force all samples into predefined classes, whereas unsupervised methods can identify new patterns but often lack clear connections to existing urban-morphological concepts. To address this challenge, this study introduces generalized category discovery into urban morphology and establishes an open-world evaluation paradigm for jointly recognizing reference morphotypes and discovering categories withheld from training. Based on this paradigm, we propose AnyMor-VLM, a novel urban-morphotype-aligned, scalable framework that leverages vision-language models capable of identifying any geometry-based urban morphotypes. The AnyMor-VLM integrates techniques of language-guided representation learning, supervised learning, self-supervised contrastive learning, and cross-modal alignment learning. Comprehensive experiments show that the AnyMor-VLM achieves the strongest overall classification performance on the proposed benchmark and improves the recognition of established reference classes and novel categories withheld from training. Two proof-of-concept applications further illustrate the use of the urban-area delineation and the identification of informal-settlement-associated fabrics. These findings provide evidence that open-world techniques can bridge knowledge-driven classification and data-driven pattern discovery in computational urban morphology.

ARTICLE HISTORY

Received 14 April 2026
Accepted 16 September 2026

KEYWORDS

Urban morphology; urban form classification; representation learning; vision-language model; semi-supervised clustering

1. Introduction

Urban form is intrinsically associated with the environmental, social, and economic performance of cities and has consequently remained a longstanding concern in urban planning and design (D'acci, Banister, and White 2024; Huang, Lu, and Sellers 2007; Zhang, Ghosh, and Park 2023). In the Conzenian tradition, urban form emerges from the relationships among the town plan, building fabric, and land and building utilization (Conzen 1960, 2004). The contemporary townscape contains forms inherited from different morphological periods and transformed through successive planning decisions and development processes. Whitehand's account of the Conzenian tradition further emphasizes morphological periods, morphological regions, the persistence of inherited forms, and the relationship between decision-making and changes in the urban landscape (Whitehand 2001). Moudon similarly identifies buildings and their associated open spaces, plots or lots, and streets as fundamental elements of urban form. She also emphasizes the importance of spatial scale and transformation over time (Moudon 1997). Urban form therefore reflects the ways in which cities organize their limited spatial resources, while being shaped by historical, institutional, functional, and socioeconomic processes (Batty 2024; Conzen 1960; Moudon 1997).

Understanding contemporary urban form remains important for evaluating established design paradigms and exploring sustainable patterns of urban development (D'acci, Banister, and White 2024; Huang, Lu, and

CONTACT Dongsheng Chen  dongsheng.chen@tum.de  Chair of Cartography and Visual Analytics, Technical University of Munich, Munich 80333, Germany

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/19475683.2026.2743305>

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Sellers 2007). However, what constitutes a sustainable urban design paradigm remains a subject of continuing debate among urban designers and planners. Examples include the garden city (Howard 1902), new urbanism (Congress for the New Urbanism 2000), and the emerging concept of 15-minute city planning (Bruno et al. 2024). A comprehensive understanding of contemporary urban forms is instrumental in facilitating the application of historical experiences and the exploration of sound, sustainable urban patterns that are aligned with the Sustainable Development Goals (Biljecki and Chow 2022; Debray et al. 2025; Huang, Lu, and Sellers 2007; Zhu et al. 2022).

At the same time, the complexity and global diversity of cities make comprehensive morphological comparisons difficult. Recent studies have consequently adopted computational representations, including street networks (Boeing 2018; W. Chen, Wu, and Biljecki 2021), urban morphometrics (Arribas-Bel and Fleischmann 2022), figure-ground maps (Moosavi 2022; J. Wang, Huang, and Biljecki 2024), block configurations (Arribas-Bel and Fleischmann 2022; D. Chen et al. 2025), etc. derived from OpenStreetMap (D. Chen et al. 2025), remote-sensing imagery (Debray et al. 2025; Zhu et al. 2022), and other data sources to support urban form analysis. Each of these representations operationalizes only part of the broader morphological system. In particular, a binary figure-ground map abstracts an urban area into the planimetric relationship between built form and open space. It retains information concerning building-footprint density, size, spacing, orientation, compactness, continuity, and spatial arrangement. Figure-ground maps therefore provide a standardized and scalable representation of geometric urban-fabric patterns.

Based on such computational representations, existing urban-form classification methods are predominantly supervised or unsupervised (W. Chen et al. 2024; Debray et al. 2025; Drici and Carpio-Pinedo 2025). Supervised approaches are guided by predefined taxonomies and domain knowledge (W. Chen, Wu, and Biljecki 2021; Zhu et al. 2022), but they generally assume that all observations belong to a fixed set of known classes approaches (Parmar et al. 2023). Previously unseen patterns may therefore be forcibly assigned to an inappropriate known category (Figure 1(a)). Unsupervised approaches avoid predefined labels and can identify recurring patterns directly from the data (J. Li et al. 2026; J. Wang, Huang, and Biljecki 2024). However, the resulting clusters are difficult to control and may not correspond to established morphotypes (Figure 1(b)). Their relationship to existing urban-morphological knowledge may consequently remain uncertain, and their interpretation can be less direct than that of predefined categories (Troisemaine et al. 2023).

The open-world learning (OWL) research offers a potential approach to addressing the myriad urban morphotypes (Z. Chen and Liu 2018). OWL research indicates that conventional supervised learning is based on the closed-world assumption, which often fails to hold in real-world scenarios, and open-world approaches are therefore necessary (Z. Chen and Liu 2018). In the open-world context, the generalized category discovery (GCD) paradigm presents a challenge: open-world models must distinguish new categories from known ones during testing or application without being retrained from scratch (Troisemaine et al. 2023). Motivated by this challenge, a growing body of research has sought to develop models with markedly stronger generalization capabilities (Troisemaine et al. 2023; Vaze et al. 2022). The strong generalizability and rich prior knowledge of large-scale pre-trained visual encoders (e.g. DINO (Vaze et al. 2022; Wen, Zhao, and Qi 2023), DINOv2 (Y. Liu, He, and Han 2025)) have shown that general-purpose visual representations can transfer effectively across a wide range of image- and pixel-level tasks. Building on these advances, a prominent direction within GCD is to jointly exploit visual and textual information through pre-trained vision-language models such as CLIP (Ouldnoughi, Kuo, and Kira 2023). Their multimodal alignment capabilities enable language-guided representations that can be effectively transferred to specialized domains (J. Cao et al. 2026; J. Wang et al. 2026). This direction is especially appealing for urban morphology, since textual descriptions of morphological attributes have long served as a key analytical device for differentiating urban form types. Taken together, inspired by the GCD paradigm, pre-trained VLM-based open-world methods hold considerable promise for extracting shared features across heterogeneous urban forms and for enabling the recognition of arbitrary urban morphotypes in open-world settings (Figure 1).

The objective of this study is to establish a scalable and typology-aligned framework for discovering geometry-based morphotypes. Our key contributions are fourfold: 1) We introduce the concept of the OWL to urban morphology and develop an open-world classification paradigm and a benchmark for urban morphology. 2) Driven by this paradigm, we propose an urban-morphotype-aligned data-driven model for identifying any urban morphotypes based on vision-language models named AnyMor-VLM.

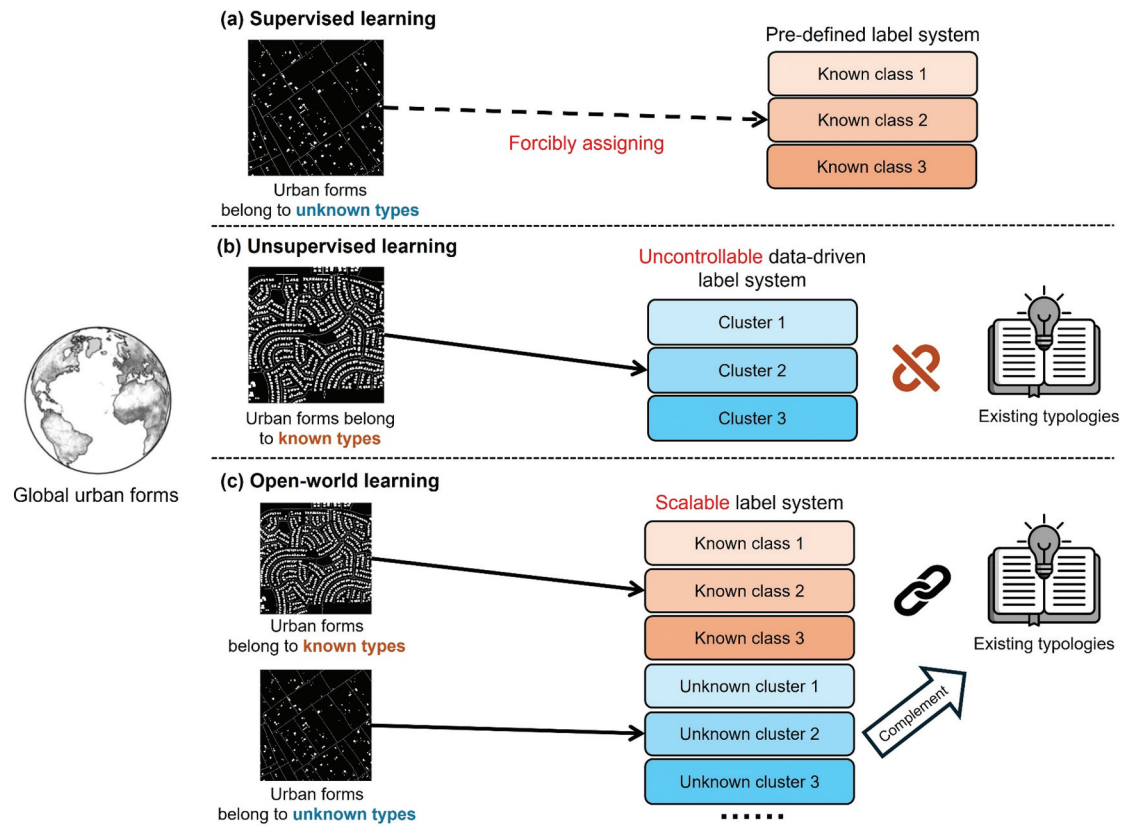


Figure 1. Challenges of existing paradigms for urban morphology classification and advance of open-world learning paradigm. (a) Supervised learning methods forcibly assign predefined classification classes to input urban forms, even if they do not belong to any predefined classes. (b) Unsupervised learning methods generate uncontrollable, data-driven categories that often lead to a disconnect from existing urban morphotypes. (c) Open-world learning methods aim to correctly classify forms belonging to known types, while clustering those belonging to unknown types into novel categories. The label system is urban-morphotype-aligned, which has the capacity to align classes with existing urban morphotypes. Furthermore, it is scalable, which allows any number of any type to be included into the system.

Methodologically, AnyMor-VLM augments state-of-the-art multi-modal GCD architectures with a dedicated map-description branch: image-captioning techniques are employed to automatically generate textual descriptions of figure-ground maps that are semantically generalizable yet expressive enough to capture salient morphological characteristics. The joint visual-textual representation is followed by a semi-supervised classification algorithm towards aligning classes with existing urban morphotypes and discovering novel types of urban morphology simultaneously. The challenge of training a classification model with strong generalizability is transformed into a problem of fine-tuning pre-trained foundation models. 3) We construct an urban morphotype benchmark containing six established urban morphotypes and evaluate models using labeled-class, unlabeled-class, novel-class, and overall accuracy under the paradigm. 4) The proposed model is further instantiated in two representative downstream tasks, i.e. urban area delineation and slum mapping, demonstrating its illustrative feasibility and transferability beyond the benchmark setting.

2. Problem formulation

This study adopts a deliberately restricted morphology-only input setting. Each sample is represented as a binary figure-ground map in which morphological elements are encoded in white and others are in black, in accordance with the findings of existing studies (Moosavi 2022; J. Wang, Huang, and Biljecki 2024). Under this representation, the model can observe geometric and layout characteristics including properties related to density, size, spacing, orientation, compactness, spatial arrangement, etc. Accordingly, this study adopts an operational definition of an urban morphotype. A figure-ground morphotype is defined as a recurring

configuration of built and open space that is geometrically distinguishable under the binary map representation. Therefore, morphological similarity in this study refers to similarity in observable figure-ground geometry. It does not imply that two urban areas are equivalent in their historical origin, land-use composition, tenure structure, street system, social function, or cultural context.

No geographic coordinates, city or country names, regional identifiers, satellite imagery, spectral information, points of interest, land-use attributes, or socioeconomic variables are provided to either the category-discovery model or the map-description generation procedure. This intentional exclusion reduces the possibility that the model relies on geographic, functional, or contextual shortcuts and enables a controlled evaluation of whether recurring built-open-space configurations can be learned and discovered from figure-ground geometry alone. The resulting inference task is therefore neither geographically conditioned nor intended to reconstruct the complete morphological identity of an urban area. Instead, it concerns the recognition of established geometry-based morphotypes and the discovery of previously unseen geometric configurations under an open-world paradigm.

Unlike existing supervised classification and unsupervised clustering paradigms for urban morphology, this paper proposes a novel, GCD-inspired classification and evaluation paradigm for open-world urban morphology. It aims to assign the task of urban morphology classification to the discovery-aware classification framework in the context of open-set classes. In the conventional GCD paradigm, the dataset D is divided into a labeled set $D_L = \{(X_i^L, y_i)\}_{i=1}^N \in \mathcal{X} \times \mathcal{Y}_L$, and an unlabeled set $D_U = \{(X_i^U, y_i)\}_{i=1}^M \in \mathcal{X} \times \mathcal{Y}_U$, where $\mathcal{Y}_L \subset \mathcal{Y}_U$. GCD requires the model to remain unaware of the labels $\mathcal{Y}_U$ in the unlabeled dataset D_U during training, but to predict $\mathcal{Y}_U$ during the testing phase. This paradigm aims to simultaneously evaluate a model's performance on both known and unknown categories. However, since this paradigm still permits models to access unlabeled data X^U during training, some methods leverage unsupervised (e.g. self-supervised) learning to maintain decent performance on the unlabeled set D_U . However, there is a possibility that they will demonstrate substandard performance in relation to data that is not contained within the dataset D . Given the nature of scarce training labels in urban morphology classification, this issue is likely to occur frequently. Thus, the conventional GCD paradigm does not fully adapt to the domain of urban morphology classification.

Our classification and evaluation paradigm of open-world urban morphology poses a more challenging requirement for the practical domain. We propose isolating a subset $D_T = \{(X_i^T, y_i)\}_{i=1}^K \in \mathcal{X} \times \mathcal{Y}_T$ as the testing set from the whole dataset D , such that $D = D_L \cup D_U \cup D_T$. The testing set D_T , including the data X^T and the labels $\mathcal{Y}_T$, are not permitted to be accessed by models during training. And $\mathcal{Y}_T$ includes three kinds of labels according to label availability (Figure 2), namely, labeled classes that are supervised trained $\mathcal{Y}_L$, unlabeled and unsupervised trained classes $\mathcal{Y}_U \setminus \mathcal{Y}_L$, and novel classes that the model has never seen or trained $\mathcal{Y}_T \setminus \mathcal{Y}_U$, such that $\mathcal{Y}_L \subset \mathcal{Y}_U \subseteq \mathcal{Y}_T$. Throughout the following formulation, the terms 'class', 'label', and 'morphotype' all refer to the operational figure-ground definition unless otherwise specified. Hence, four accuracies can be calculated, including the overall accuracy OA for the whole D_T , the labeled-class accuracy ACC_L , the unlabeled-class accuracy ACC_U , and the novel-class accuracy ACC_{novel} .

The proposed paradigm consists of two phases. (1) In the model development phase, D_T must include the data with novel classes, such that $\mathcal{Y}_U \subset \mathcal{Y}_T$ (Figure 2(a)). During testing, the model must be tested on all kinds of labels $\mathcal{Y}_T$, thereby evaluating the capabilities of the model architecture and parameter settings on classifying the known classes and generalizing the unknown classes. (2) During the final model training phase, to fully leverage the limited data for training a model capable of classifying arbitrary urban morphotypes, we redistribute the labeled set D_{L-full} and the unlabeled set D_{U-full} in the dataset D . D_{L-full} is maximized to fully utilize existing labels, while D_{U-full} retains the minimal size of $\mathcal{Y}_U \setminus \mathcal{Y}_L$ as 1 (Figure 2(b)). This ensures that its unknown-class accuracy serves as an indicator for adjusting final model parameters. Here, $\mathcal{Y}_{L-full} \subset \mathcal{Y}_{U-full} = \mathcal{Y}_T$. The redistributed D_{L-full} and D_{U-full} are used to retrain a final model for deployment.

3. Methodology

3.1 Overview of the urban-morphotype-aligned, scalable approach

Figure 3(a) illustrates the workflow for applying the proposed AnyMor-VLM framework to achieve open-world urban morphology classification. During the first step, AnyMor-VLM is first fine-tuned (See Sections 3.2

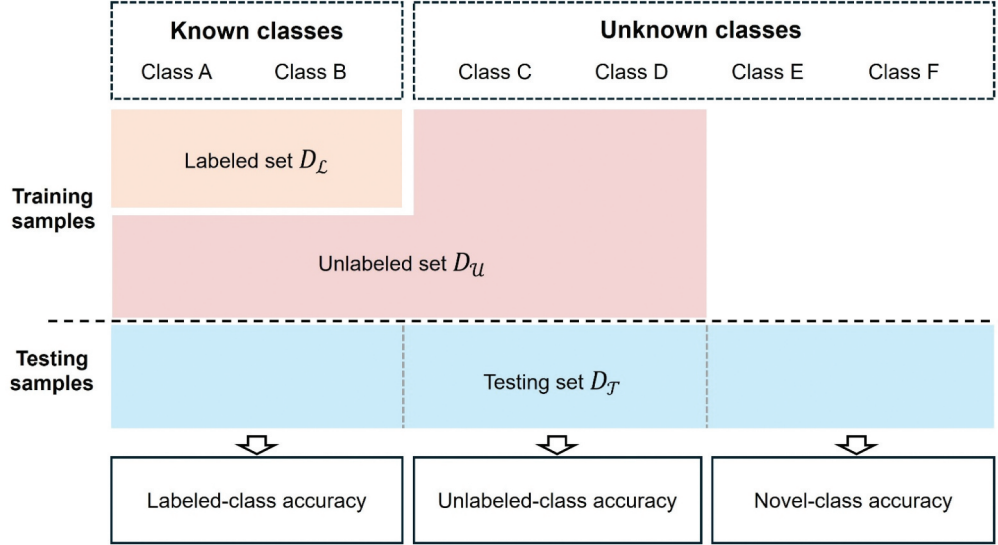
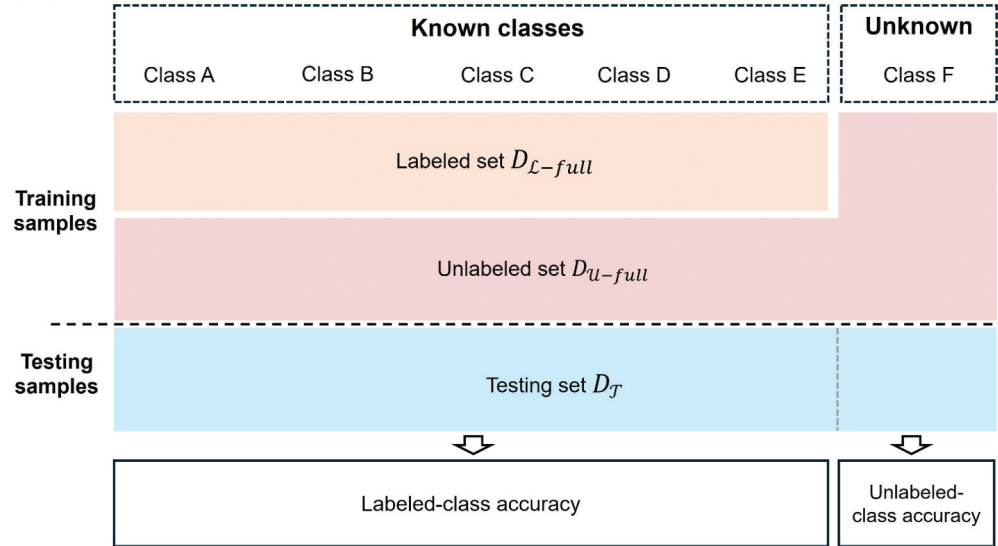
(a) Model development phase**(b) Final model deployment phase**

Figure 2. Diagram of the proposed evaluation paradigm for open-world urban morphology classification.

and 3.3 for more details) and then is deployed to generate morphological representations for the input figure-ground maps. In the second step, the generated morphological representations are projected on the embedding space. The semi-supervised classification is applied to identify the known morphotypes and unknown morphotypes (See Section 3.4).

3.2 Architecture of the AnyMor-VLM

Figure 3(b) delineates the details of the architecture of the proposed AnyMor-VLM framework. It consists of three VLM-based branches to extract comprehensive morphological representations from figure-ground maps of urban form. CLIP ViT-B/16, the widely used multi-modal backbone (He, Liu, and Han 2025), is selected as the VLM backbone of three branches, serving not only as the image feature extractor but also as the shared alignment interface among all three branches because it has a strong common representation space for visual information and textual information (Ouldnooghi, Kuo, and Kira 2023). A visual-only encoder such as DINOv2 could provide strong structural representations and has demonstrated effectiveness in earth observation and remote sensing tasks (Osco et al. 2023; R.; E. Wang et al. 2025), but it is not directly

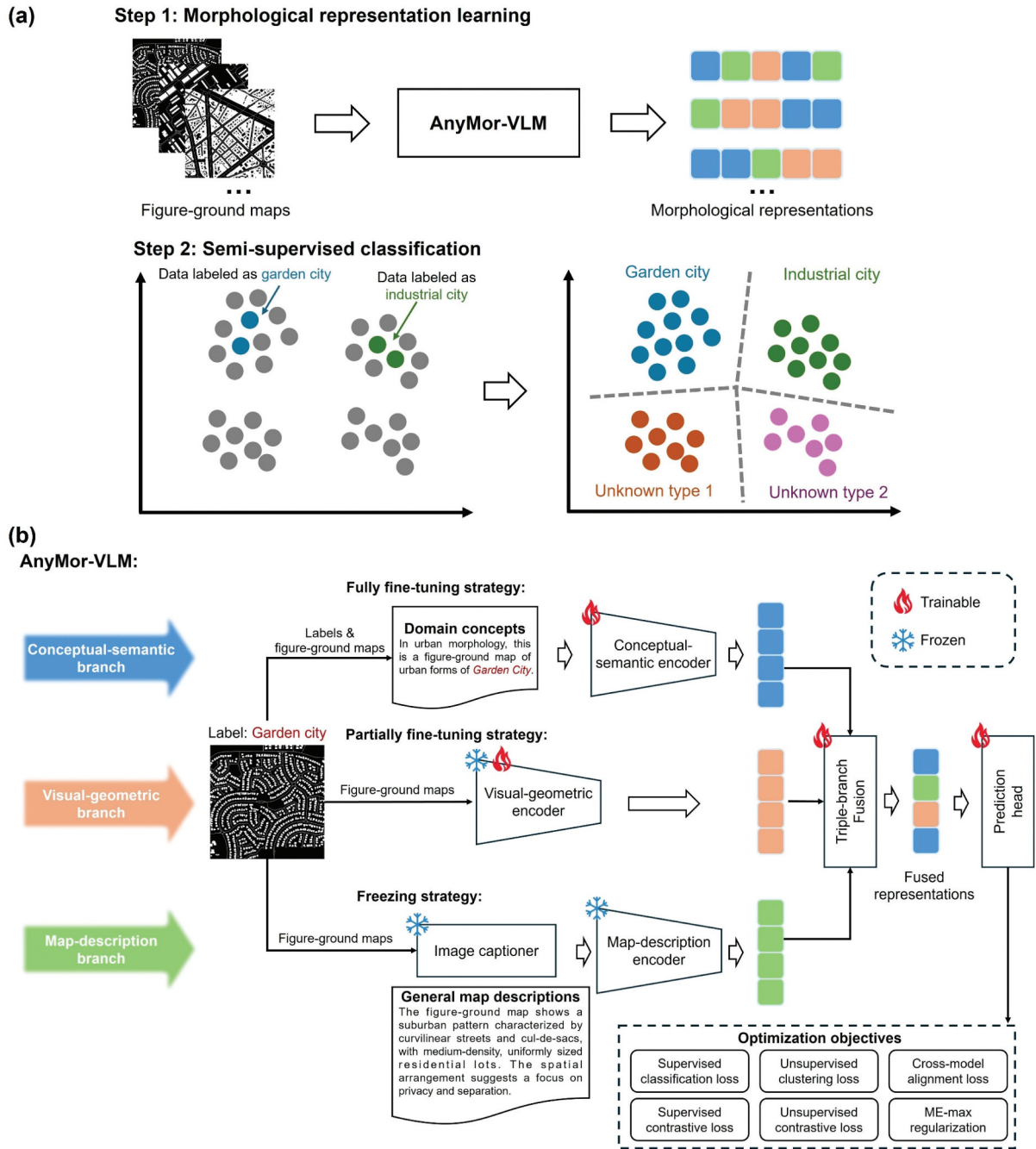


Figure 3. (a) Workflow towards the open-world urban morphology classification in this study. (b) The architecture of the proposed AnyMor-VLM model.

interchangeable within the current aligned visual-text architecture. Three distinct fine-tuning strategies are adopted to make full use of the original generalizability and the domain transferability of the VLMs. Furthermore, a fusion projector integrates the three generated characteristics as the final fused representation, in order to make full use of the existing information. The specific details of them are shown below.

- **Conceptual-semantic branch** has a core conceptual-semantic encoder that takes figure-ground maps and morphotype labels as input and generates label information for any unlabeled figure-ground maps (Figure 3(b) Blue branch). The conceptual-semantic encoder aims to establish a linear mapping relationship between the visual features of the figure-ground map and its conceptual semantics via natural language-guided representation learning. In this study, we apply a fully fine-tuned CLIP model $f_{CLIP}(\cdot)$

and a simple fully connected projector g to jointly form the conceptual-semantic encoder module $f_{CS} = g(f_{CLIP}(\cdot))$. Firstly, the known labels are integrated with a prompt template '*In urban morphology, this is a figure-ground diagram of urban forms of {Type} with {Characteristic}*' to construct the concept of the known morphotypes. Secondly, the representations with label information are generated by encoding the concept with the CLIP textual encoder. Next, the CLIP model $f_{CLIP}(\cdot)$ is fully fine-tuned via model distillation and modal alignment techniques to learn the link between maps and concepts while ensuring it remains within the representation space of the original CLIP model (E. Wang et al. 2025). Finally, the CLIP model enables the generation of conceptual semantic representations for figure-ground maps.

- **Visual-geometric branch** has a visual-geometric encoder that takes only figure-ground maps as input and extracts the geometric information from the figure-ground maps (Figure 3(b) Orange branch). The visual-geometric encoder is built on a CLIP visual encoder and a partially fine-tuning strategy was employed, entailing the training of only the final layer of the encoder. The objective is to optimize the CLIP model's retention of its original generalizability while concurrently transferring the capacity to recognize visual-geometric features from figure-ground maps of urban morphology. In summary, the visual-geometric branch can be formulated as $f_{VG} = f_{CLIP-Vis}(\cdot)$.
- **Map-description branch** is a module that takes only a figure-ground map as input and generates information of map-related textual features describing figure-ground maps (Figure 3(b) Green branch). This branch comprises two components. 1) Image captioner $f_{Caption}(\cdot)$ is a VLM that generates textual descriptions of images. Image captioning has emerged as a prominent research area in computer vision, with an expanding portfolio of models proposed in recent years (Stefanini et al. 2023). In our framework, the captioning module is used to generate natural-language descriptions of the spatial patterns exhibited by figure-ground maps, which we refer to as 'map descriptions'. To select an appropriate captioner, we systematically compared several representative models, including BLIP, BLIP-2, and GPT-4o (Vision). GPT-4o was retained because, unlike BLIP/BLIP-2, it produced morphology-oriented vocabulary (e.g. *orthogonal grid*, *cul-de-sac*, *irregular street pattern*) with substantially fewer hallucinated objects. A detailed account of the candidate models (including versions, hyperparameter settings, prompt design, and qualitative and quantitative comparisons of their outputs) is provided in the *Selection of candidate image captioners* section of the Supplementary Materials (Table S4; Figure S2). To ensure that this design choice does not compromise reproducibility, we further conducted a controlled repeated-sampling experiment that quantifies the stability of GPT-4o-generated map descriptions under identical inputs (see *Reproducibility of GPT-4o map descriptions* section and Table S5 in the Supplementary Materials). The results show that the captions generated remain highly consistent across repeated queries, confirming that GPT-4o provides a stable and reliable captioning backbone within our pipeline (Figure S3). 2) Textual encoder $f_{CLIP-Txt}(\cdot)$: encodes the generated map descriptions using the frozen CLIP textual encoder and maps them to the embedding space consistent with the visual features. In short, the map-description branch can be formulated as $f_{MD} = f_{CLIP-Txt}(f_{Caption}(\cdot))$.
- **Triple-branch fusion module** is designed to integrate the characteristics from three branches. Multimodal fusion methods can broadly be categorized as feature-level fusion, interaction-based fusion, and decision-level fusion (Ateş and Yurtsever 2026). In this paper, the proposed fusion module follows a lightweight feature-level strategy, i.e., a simple nonlinear fully connected projection layer g . It is trainable, enabling the model to select the characteristics from the three branches based on feedback from training data during the training process. This design is appropriate because all three branches are encoded within the same pretrained CLIP embedding space and produce fixed-length global representations with matched dimensionality. Consequently, the model does not need to learn alignment between heterogeneous feature spaces or fine-grained token-level correspondences. More complex mechanisms (Ateş and Yurtsever 2026), such as cross-attention, bilinear fusion, or mixture-of-experts routing, would introduce substantially greater model capacity and may increase the risk of overfitting given the scale of the current benchmark. This module can be formulated as $f_{Fusion} = g(f_{CS}(\cdot), f_{VG}(\cdot), f_{MD}(\cdot))$.

In the step of model fine-tuning, the prediction head first employs a parametric cluster layer (DINO-head) (E. Wang et al. 2025) to facilitate model optimization (Step 1 in Figure 3(a)). While in the step of deploying the final model, a semi-supervised K-Means++ algorithm replaces the parametric clustering layer (Step 2 in Figure 3(a)). This modification is made to address the challenges posed by scalable data samples and novel classes. The non-parametric clustering method is characterized by its capacity to adapt to the number of clusters without the necessity of re-training, which aligns with the requirement of the step of deploying final model.

3.3 Urban-morphotype-aligned strategy for fine tuning

We implement a fine-tuning strategy that simultaneously enhances the capability of learning morphological representations and aligning the representations with existing morphotypes. It aims to transfer the original generalizability of AnyMor-VLM to urban morphology domains. This strategy achieves the urban-morphotype-aligned capability by using supervised learning objectives while enhancing the morphological representation learning by using unsupervised learning objectives, as illustrated in the optimization objectives of Figure 3(b). Furthermore, the vision-language alignment learning objective is added into the strategy to integrate the trainable conceptual-semantic and visual-geometric representations. Last but not least, mean-entropy maximization regularization is integrated into the multiple objectives to avoid degenerate predictions.

For any module mentioned above, the input data x_i from any mini-batch B is processed into its corresponding representation z_i and the probability distribution p_i of the prediction results. Supervised learning is comprised of two components: traditional supervised classification, which targets the predicted probability distribution, and contrastive learning, which targets the representations. The classification component employs the cross-entropy loss $H(\cdot)$, the most commonly used loss function for multi-class classification problems. The loss function for supervised learning can be formulated as follows.

$$L_{s-cls} = \frac{1}{|B^{\mathcal{L}}|} \sum_{i \in B^{\mathcal{L}}} H(y_i || \sigma(p_i/\tau)) \quad (1)$$

where $\sigma(\cdot)$ refers to the softmax function and $H(y_i || \sigma(p_i, \tau))$ indicates the cross-entropy term between the ground-truth label y_i and the prediction result $\sigma(p_i, \tau)$. $B^{\mathcal{L}}$ and τ represent the labeled subset of B and the temperature. Supervised contrastive learning aims to minimize the distance between the representations z_i and z_p of data with the same label (Khosla et al. 2020). Its loss function is formulated as:

$$L_{s-con} = -\frac{1}{|B^{\mathcal{L}}|} \sum_{i \in B^{\mathcal{L}}} \frac{1}{|P_i|} \sum_{p \in P_i} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in B^{\mathcal{L}}, a \neq i} \exp(z_i \cdot z_a/\tau)} \quad (2)$$

where P_i refers to the set of the samples with the same label except sample i . z_a indicates any representations except sample i in $B^{\mathcal{L}}$. In summary, the loss function of supervised learning is stated as $L_{super} = L_{s-cls} + L_{s-con}$.

Unsupervised learning involves the random cropping of two views, x_i and x_i' , from any given figure-ground map. This process yields corresponding representations, z_i and z_i' , as well as the corresponding probability distributions for the prediction results, p_i and p_i' . Subsequently, it involves two kinds of techniques, including unsupervised clustering learning and unsupervised contrastive learning. Unsupervised clustering learning clusters the results of all two-view data in mini-batch B based on their probability distributions (Wen, Zhao, and Qi 2023) and employs a self-distillation loss, as follows.

$$L_{u-cls} = \frac{1}{|B|} \sum_{i \in B} H(\sigma(p_i'/\tau) || \sigma(p_i/\tau)) \quad (3)$$

While the unsupervised contrastive learning employs a self-supervised strategy to minimize the distance between the representations of two views (X. Liu et al. 2023), with its loss function stated as:

$$L_{u-con} = -\frac{1}{|B|} \sum_{i \in B} \log \frac{\exp(z_i \cdot z'_i / \tau)}{\sum_{a \neq i, a \in B} \exp(z_i \cdot z'_a / \tau)} \quad (4)$$

Furthermore, a mean-entropy maximization regularization term $H(\bar{p}||\bar{p})$ is used for unsupervised learning to avoid degenerate predictions, such as cluster collapse, imbalance, or constant outputs (Assran et al. 2022). $\bar{p} = \frac{1}{2|B|} \sum_{i \in B} (\sigma(p'_i / \tau) + \sigma(p_i / \tau))$ indicates the mean softmax probability of the two-view data's probability distributions. In summary, the unsupervised learning loss function $L_{unsuper} = L_{u-cls} + L_{u-con} - \varepsilon H(\bar{p})$, where ε is the weight of the mean-entropy maximization regularization term and is usually set to 1.

The trainable conceptual-semantic branch $f_{CS}(\cdot)$, the visual-geometric branch $f_{VG}(\cdot)$, and the fusion module of the triple branches $f_{Fusion}(\cdot)$, all employ corresponding loss functions from supervised and unsupervised learning. They can be formulated as a set $\{L^{CS} = \lambda L_{super}^{CS} + (1 - \lambda) L_{sunuper}^{CS}, L^{VG} = \lambda L_{super}^{VG} + (1 - \lambda) L_{sunuper}^{VG}, L^{Fusion} = \lambda L_{super}^{Fusion} + (1 - \lambda) L_{sunuper}^{Fusion}\}$, where λ is the supervision weight.

In addition, cross-modal alignment loss function L_{cross} is set to keep consistency between the trainable conceptual-semantic branch f_{CS} and visual-geometric branch f_{VG} (E. Wang et al. 2025). The Kullback–Leibler divergence D_{KL} is used to calculate mutual information for the two branches.

$$L_{cross} = \frac{1}{2|B|} \sum_{i \in B} (D_{KL}(\sigma(z_i^{CS} \cdot \mathcal{P}^{VG}) || \sigma(z_i^{VG} \cdot \mathcal{P}^{CS})) + D_{KL}(\sigma(z_i^{VG} \cdot \mathcal{P}^{CS}) || \sigma(z_i^{CS} \cdot \mathcal{P}^{VG}))) \quad (5)$$

where $\mathcal{P}$ represents the prototypes of all the categories with labels. Here, the prototypes are the mean representations of the samples with the same labels.

In summary, the overall loss function L consists of the loss function of conceptual-semantic branch L^{CS} , visual-geometric branch L^{VG} , fusion of the triple branches L^{Fusion} as well as the cross-modal alignment loss L_{cross} , i.e. $L = L^{CS} + L^{VG} + \lambda_{fusion} L^{Fusion} + L_{cross}$. The weight λ_{fusion} for L^{Fusion} is set to 2 according to the result of parameter sensitivity analysis shown in Figure S4.

3.4 Semi-supervised K-Means++ for scalable label systems

During the final model deployment phase, the corresponding morphological representation can be calculated for any new input figure-ground map from the world via the fine-tuned AnyMor-VLM. These new representations from the world are then compared with the existing representations of known urban morphotypes in the dataset D .

The semi-supervised K-Means++ algorithm is used to compare the new representations with the existing representations. It initially computes the labeled centres of clusters from the representations of the dataset D with the labels of known morphotypes, while randomly generates unlabeled centres for unknown morphotypes. And then the cluster centres are iteratively updated in a manner of the traditional K-Means algorithm. Once the algorithm converges, the new representations are categorized into known morphotypes when they align more closely with the labeled centres, or into a novel morphotype when they maintain distance from all the labeled centres (Vaze et al. 2022).

Determining the number of clusters. This study uses the combination of cluster accuracy and visual analytics to determine the value of K in the final model deployment phase. Due to the absence of labels in all the new representations, it is challenging to determine K for the semi-supervised K-Means++ algorithm. In this case, the solution is twofold. 1) In accordance with the GCD paradigm (Vaze et al. 2022), the optimal K is determined by evaluating the clustering accuracy for the dataset D with the labels denoting known morphotypes. This approach guarantees the accurate classification of data characterized by known morphotypes. 2) We propose using visual analytics techniques, i.e. dimensionality reduction methods to determine K . This study applies the method of Uniform Manifold Approximation and Projection (UMAP) to project the morphological representations into a two-dimensional embedding space (Becht et al. 2019). Subsequently, the number of clusters can be visually observed in the two-dimensional embedding space. The final determination of K is contingent upon the consideration of these two results.

3.5 Accuracy assessment

In accordance with the GCD paradigm (Vaze et al. 2022), classification accuracy is evaluated using a clustering accuracy via the Hungarian algorithm. As a well-established combinatorial optimization procedure, the Hungarian algorithm determines the optimal one-to-one assignment between predicted cluster indices and ground-truth class labels, thereby resolving the label-permutation ambiguity inherent in unsupervised and semi-supervised clustering. Given the ground truth labels y_i and the model's predictions $\hat{y}_i$, the clustering accuracy is defined as follows.

$$ACC = \frac{1}{|D_T|} \sum_{i \in D_T} (y_i = p(\hat{y}_i)) \quad (6)$$

where p is the optimal permutation that matches the $\hat{y}$ to y in the clustering context.

To provide a more comprehensive evaluation, we further adopt two metrics that are widely used in the clustering research, i.e. the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). ARI assesses clustering quality by considering all pairs of samples and comparing their relative relationships under the ground-truth partition and the predicted partition (Chacón and Rastrojo 2023). For any pair of samples, four mutually exclusive cases can be distinguished:

- SS (Same–Same): the two samples belong to the same class in the ground truth and are also assigned to the same cluster in the prediction (correct agreement).
- DD (Different–Different): the two samples belong to different classes in the ground truth and are assigned to different clusters in the prediction (correct disagreement).
- SD (Same–Different): the two samples share the same ground-truth class but are assigned to different predicted clusters (incorrect disagreement).
- DS (Different–Same): the two samples belong to different ground-truth classes but are assigned to the same predicted cluster (incorrect agreement).

Based on these four cases, the Rand Index is defined as:

$$RI = \frac{SS + DD}{SS + SD + DS + DD} \quad (7)$$

To ensure that the expected score under random label assignment is centred at zero, ARI corrects the Rand Index for chance by subtracting its expectation and normalizing by the maximum attainable value:

$$ARI = \frac{RI - Expected(RI)}{Max(RI) - Expected(RI)} \quad (8)$$

ARI takes values in $[-1,1]$, where 1 denotes a perfect match with the ground truth, 0 corresponds to a partition no better than random assignment, and negative values indicate agreement worse than expected by chance.

NMI, in contrast, is grounded in information theory and quantifies the reduction in uncertainty about the predicted labels once the ground-truth labels are known (and vice versa) (K. Cao, Brbic, and Leskovec 2022). Its equation involves two fundamental quantities:

- Mutual Information $I(\mathbf{y}, \hat{\mathbf{y}})$: the amount of information shared between the ground truth labels $\mathbf{y}$ and the predicted labels $\hat{\mathbf{y}}$.
- Entropy $H(\mathbf{y})$, $H(\hat{\mathbf{y}})$: the uncertainty inherent of the ground-truth and predicted label distributions, respectively.

The equation for NMI is as follows:

$$NMI(\mathbf{y}, \hat{\mathbf{y}}) = \frac{2 * I(\mathbf{y}, \hat{\mathbf{y}})}{\sqrt{H(\mathbf{y}) * H(\hat{\mathbf{y}})}} \quad (9)$$

NMI ranges from 0 to 1, with 1 indicating that the predicted partition perfectly recovers the ground-truth partition and 0 indicating that the two labelings are statistically independent.

ARI and NMI provide extra complementary perspectives on the evaluation of open-world urban morphology classification. ARI is robust against inflated scores caused by random noise and is therefore well suited for verifying that improvements are not attributable to chance, whereas NMI is more sensitive to clustering quality when class sizes are highly imbalanced. Using the Hungarian-matched clustering accuracy, ARI, and NMI in conjunction thus enables a more balanced and reliable assessment of the proposed method.

4. Experiments and analyses

4.1 Experimental setup

- Data collection and Benchmark

The way of data collection for figure-ground maps involves the extraction of urban morphological elements (building footprints and streets) from the open-source OpenStreetMap. These features are then converted into binary images, with white representing the morphological elements and black representing the other. Each figure-ground map has been set to a resolution of 400×400 metres, in accordance with the setting delineated in (Fleischmann, Feliciotti, and Kerr 2022). The $400 \text{ m} \times 400 \text{ m}$ observation window is an experimental sampling unit used to standardize spatial extent, image dimensions, and visual resolution across the benchmark, which does not correspond to an internally homogeneous morphological region.

Guided by the proposed paradigm, we also establish an operational benchmark to evaluate category discovery within the operational reference system. In accordance with the subdivision in urban morphotypes across historical periods, this study sets six well-established types of urban form as the initial morphotypes as well as the ground-truth labels (Figure 4), i.e. ‘pre-industrial city’, ‘industrial city’, ‘garden city’, ‘modernist planning’, ‘neo-traditional planning’, and ‘informal settlements’. The corresponding representative areas for these six urban morphotypes are adopted from the 42 cases collected worldwide by Fleischmann, Feliciotti, and Kerr (2022), which provides a methodological foundation for systematic of urban form (Table S1). However, the six morphotypes should be understood as a historically informed and operational reference typology. They provide recognizable reference patterns for evaluating whether a category-discovery model can retain established morphological knowledge while identifying categories withheld from training (Table S2). Some categories originate from urban design and planning traditions that were initially formulated primarily in European and North American contexts, whereas all of them refer to recurrent geometry patterns observed across different geographical settings. They therefore do not necessarily share the same historical or conceptual origin.

To expand the sample size, 100 points for each case were randomly sampled within a 5-kilometre radius of the 42 representative areas (Figure S1). Subsequently, these samples were subjected to a screening process by domain experts through visual interpretation, with the objective of selecting those that exhibited analogous morphological characteristics. The details of the data-screening procedure can be found in *Section S1 Details of data labeling* in the Supplementary Materials. After inter-rater reliability testing (Table S3; Appendix A), 2,414 samples with 6 urban morphotypes ultimately resulted in the construction of an urban morphotype benchmark. The number of samples per morphotype ranged from 336 to 459. The benchmark labels have heterogeneous conceptual provenance (Table 1). To avoid treating planning paradigms, historical formations, and settlement conditions as ontologically equivalent, each morphotype is operationalized through its observable figure-ground manifestation.

In the model development phase, the urban morphotype benchmark is partitioned into three subsets, i.e. D_L , D_U , D_T . To emulate realistic conditions in which known morphotypes are typically outnumbered by unknown ones, we adopt a class-ratio configuration of 2:2:2 across the three subsets. Under this setting, the model can only perceive the labels from the 2 classes as the known morphotypes, whereas the other four classes remain unknown throughout training. The assignment of morphotypes to the 2:2:2 partition is generated at random and repeated ten times to mitigate the influence of any particular class configuration on model performance. Identical random assignments are applied consistently across all competing models to ensure a fair comparison.

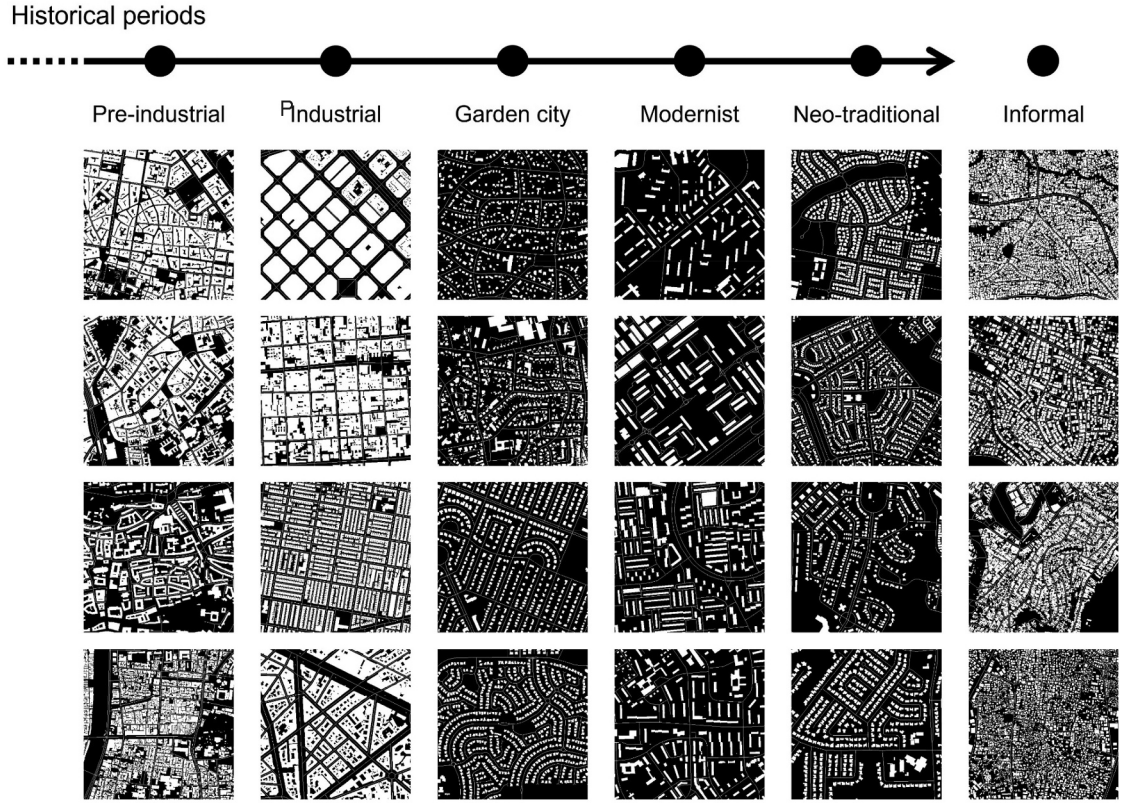


Figure 4. Six urban morphotypes across historical periods. Four representative examples in the urban morphotype benchmark are illustrated for each morphotype.

In the final deployment phase, all six morphotypes were used to train the full model, yielding a cross-regional proof-of-concept classifier based on the AnyMor-VLM framework. To assess its behavior in practical settings, we further collected unseen figure-ground maps from four selected regions, i.e. Barcelona (Spain), Singapore, Stratton and its surroundings (Australia), and Nairobi (Kenya).

- Implementation details

The dataset D was partitioned into an 80% training set and a 20% testing set (D_T). The training set is further comprised of two sets, a 40% labeled set D_L , and a 60% unlabeled set D_U . Figure-ground maps were treated as image inputs, with data augmentation techniques including random cropping and flipping applied. For the CLIP model, we employed ViT-B/16. The input text for the CLIP textual encoders is limited to 30 tokens. In the conceptual-semantic branch, the CLIP model outputs 512-dimensional semantic representations. Consequently, the subsequent linear fully connected projector is configured to accept a 512-dimensional input and produce a 768-dimensional output. In the visual-geometric branch, the CLIP visual encoder outputs a hidden layer with a dimension of 768. In the map-description branch, the text encoder outputs a dimension of 512. In the fusion module, the input dimension of the nonlinear fully connected projector is set to $[768, 768, 512]$ to sequentially match the output dimensions of the three branches, with an output dimension of 768. Finally, the input dimension for the parametric clustering layer (DINO-Head) used in the model development phase is 768. It incorporates a three-layer MLP with a fixed output dimension of 6 to match the six urban morphotypes in the benchmark. The temperature values for the loss functions are as follows: 1.0 for self-supervised contrastive learning; 0.07 for supervised contrastive learning; 0.1 for supervised classification; 0.07 for unsupervised clustering; and 0.1 for the mean-entropy maximization regularization term. The stochastic gradient descent optimizer is used for the joint fine-tuning process with a learning rate of 0.01 and a decay weight of 0.00001. All models converged within 70 epochs. The AnyMor-VLM model is trained via Intel i9-13900K, and NVIDIA GeForce RTX 4080 (total memory of 16375MB, 76 multi-processors).

Table 1. Operational definition of the six urban morphotypes.

Label name	Benchmark reference morphotype		Conceptual provenance		Operational figure-ground characteristics		Not inferred from the morphotype	
	Pre-industrial	Industrial	Historical epoch and socio-technological stage	Techno-economic paradigm	dense, compact pattern; labyrinthine organic layout	Actual age of buildings, historical significance, cultural heritage status, or original social functions.	Actual age of buildings, historical significance, cultural heritage status, or original social functions.	Actual age of buildings, historical significance, cultural heritage status, or original social functions.
Garden city		Industrial city-like form	Techno-economic paradigm	Planning and design paradigm	dense compact form; regular grid or linear pattern	Current economic use, pollution levels, industrial ownership.		
Modernist	Garden-city-derived form		Planning and design paradigm		low-density form; curvilinear organic pattern	Garden city ideology, planning intention or historical origin		
Neo-traditional	Modernist planning-derived form		Planning/design and architectural history		Geometric pattern, elongated buildings, low density	Adherence to modernist ideology or actual construction period		
Informal	New urbanism-derived form		Planning and design paradigm		partially compact form & modified grid	New urbanism ideology, planning intention or historical origin		
	Informal-settlement reference form		Settlement condition and urbanization process		High density, small buildings, irregular pattern	Informality, legality, poverty, tenure or socioeconomic exclusion		

GPU memory consumption is 16 GB. It processes 24.88 figure-ground maps per second during the model training stage, while processing 19.32 figure-ground maps per second during the model inference stage. GPT-4o was not called during AnyMor-VLM training or inference. The GPT-4o stage was performed once as an offline preprocessing step, using one API request per map and a total of 2,414 requests. Caption generation required an average response time of 0.5 s per map, and incurred an API cost of approximately \$0.003825 per map. The generated descriptions were cached and reused across all model configurations, random seeds, and training runs.

- Baselines

To comprehensively evaluate the proposed AnyMor-VLM, we benchmark it against seven baselines spanning three methodological categories, namely the random permutation baseline, baselines from urban morphology, and baselines from GCD.

(1) Random permutation tests are conducted to establish the empirical distribution of random label assignments and serve as the most fundamental baseline. The class configuration, test sample distribution, and random seed are held identical to those of the proposed model, thereby ensuring an equivalent input setting. The predictions of the proposed model are then randomly permuted, and this permutation procedure is repeated 1,000 times for each input configuration. The mean and standard deviation of the resulting evaluation metrics are reported as the chance-level reference.

(2) Regarding the baselines from the field of urban morphology, we select two classic methods for urban form clustering, including urban morphometrics and autoencoder, as follows.

Urban morphometrics constitutes a widely adopted quantitative approach to the study of urban form (Oliveira and Porta 2025). Through a set of theory-driven morphological indicators, it enables the characterization of urban form across scales, from overall spatial patterns to fine-grained local details. Recent studies have demonstrated its effectiveness in distinguishing urban forms across large geographical extents and diverse historical layers (Arribas-Bel and Fleischmann 2022; Fleischmann et al. 2022, 2026). This baseline adopts an advanced morphometric set specifically designed for cross-national clustering of urban form (Fleischmann et al. 2026), comprising 59 indicators that jointly capture buildings, streets, and enclosed tessellation cells at small, medium, and large scales. The indicators are standardized and treated as feature representations of the input data. These representations replace the AnyMor-VLM morphological representations illustrated in Figure 3(a) and are subsequently fed into the semi-supervised classification pipeline, thereby ensuring methodological comparability across baselines.

Autoencoder represents a foundational approach in representation learning for urban form (Moosavi 2022), and pioneered the automated study of urban form at the global scale. This baseline replicates the original Autoencoder architecture, which comprises five convolutional encoder layers with 15, 15, 15, 10, and 10 kernels of size 5×5 , each followed by a ReLU nonlinearity, together with five deconvolutional decoder layers of symmetric configuration. Two fully connected layers bridge the encoder and decoder to the latent representation. The Autoencoder is trained on $D_{\mathcal{L}} \cup D_u$, and the learned representations are subsequently input into the semi-supervised classification module for comparison with the other methods.

(3) Regarding the baselines from the field of GCD, we collect one baseline and three state-of-the-art under the common CLIP representation framework from the GCD literature. This controlled design isolates differences arising from the category-discovery strategy and from the proposed map-description branch.

Pre-trained CLIP serves as the base method, integrating the pre-trained CLIP visual encoder for representation extraction from figure-ground maps with a semi-supervised K-Means++ algorithm for classification (Vaze et al. 2022). This baseline reflects the intrinsic representational capacity of the CLIP model.

GCD-CLIP originates from the seminal GCD proposed by Vaze et al. (2022) and is subsequently adapted with the CLIP backbone (E. Wang et al. 2025). It applies supervised and self-supervised contrastive learning to fine-tune the CLIP representation space, followed by semi-supervised K-Means++ for classification.

SimGCD-CLIP constitutes an enhanced variant of GCD-CLIP (Wen, Zhao, and Qi 2023). While it likewise employs supervised and self-supervised contrastive learning based on CLIP, it dispenses with the semi-supervised K-Means++ step and instead introduces a parametric clustering layer to jointly guide the fine-tuning of the CLIP backbone.

GET is a pioneering model that leverages semantic information encoded in category names (E. Wang et al. 2025). It comprises two branches: a visual branch and a semantic branch. The semantic branch exploits category names from the labeled data and generates pseudo-label semantics directly from images, while the visual branch fine-tunes the final module of the CLIP visual encoder. Building upon SimGCD-CLIP, GET further incorporates supervised classification learning, unsupervised clustering learning, and cross-modal alignment learning into a unified training strategy.

4.2 Model performance

We benchmarked AnyMor-VLM against a set of baselines and state-of-the-art methods, with the aggregate results summarized in Table 2. Four principal observations emerge from this comparison.

- (1) Traditional expert-designed morphological indicators remain an effective baseline in the novel evaluation paradigm. Among the baselines from the field of urban morphology, the classical urban morphometrics approach yields solid and stable performance, improving OA by 24.05% points, ARI by 0.216, and NMI by 0.287 relative to a random-permutation reference. In contrast, the autoencoder is markedly less competitive. Although it exhibits low standard deviations across the three evaluation metrics, its mean performance lags substantially behind urban morphometrics. This contrast underscores the robust generalization afforded by hand-crafted morphological descriptors.
- (2) Pre-trained CLIP offers competitive general-purpose representations. Under identical experimental settings, the off-the-shelf representations produced by pre-trained CLIP are on par with the strongest baseline from urban morphology. While its OA is roughly on par with that of urban morphometrics, CLIP surpasses the latter in both ARI and NMI. This suggests that pre-trained CLIP models already encode features that discriminate among urban morphotypes to a non-trivial extent, and that large foundation models constitute a promising avenue for quantitative urban morphology research.
- (3) The GCD technique substantially advances the open-world urban morphology classification. Although originally developed for natural-image recognition, the three GCD methods evaluated here stably outperform both the urban-morphology baselines and pre-trained CLIP across supervised,

Table 2. Performance comparison of AnyMor-VLM with the baselines and state-of-the-art methods under the 2:2:2 setting in the model development phase. Means and standard deviations are computed over 10 random 2:2:2 configurations. Bold values denote the best performance, and underlined values indicate the second-best.

Methods	OA%	ACC _C %	ACC _{UL} %	ACC _{novel} %
Random	21.56 ± 0.95	22.49 ± 5.17	20.34 ± 5.94	21.38 ± 5.48
Morphometrics	45.61 ± 2.49	61.35 ± 12.73	40.31 ± 18.78	33.38 ± 17.75
Autoencoder	35.55 ± 0.76	38.91 ± 7.15	31.58 ± 13.28	35.28 ± 6.64
Pre-trained CLIP	45.34 ± 2.08	45.37 ± 12.92	51.68 ± 18.85	39.47 ± 17.60
GCD-CLIP	62.77 ± 6.32	75.11 ± 17.66	67.66 ± 18.89	43.13 ± 17.99
SimGCD-CLIP	61.24 ± 5.65	68.19 ± 15.08	64.67 ± 22.10	49.07 ± 15.78
GET	62.34 ± 2.85	77.78 ± 14.91	66.32 ± 20.71	41.47 ± 17.22
AnyMor-VLM (ours)	64.87 ± 3.62	74.05 ± 12.56	70.67 ± 21.75	<u>48.35 ± 16.10</u>
	ARI _{overall}	ARI _C	ARI _{UL}	ARI _{novel}
Random	0 ± 0.003	0 ± 0.007	0 ± 0.007	0 ± 0.007
Morphometrics	0.216 ± 0.023	0.370 ± 0.153	0.305 ± 0.210	0.214 ± 0.128
Autoencoder	0.120 ± 0.007	0.155 ± 0.107	0.124 ± 0.070	0.116 ± 0.080
Pre-trained CLIP	0.246 ± 0.019	0.291 ± 0.067	0.345 ± 0.129	0.218 ± 0.113
GCD-CLIP	0.427 ± 0.080	0.640 ± 0.227	0.505 ± 0.184	<u>0.308 ± 0.134</u>
SimGCD-CLIP	0.437 ± 0.065	0.610 ± 0.057	<u>0.594 ± 0.154</u>	0.276 ± 0.165
GET	0.441 ± 0.041	0.693 ± 0.056	<u>0.546 ± 0.167</u>	0.317 ± 0.177
AnyMor-VLM (ours)	0.465 ± 0.055	0.613 ± 0.101	0.663 ± 0.172	0.296 ± 0.191
	NMI _{overall}	NMI _C	NMI _{UL}	NMI _{novel}
Random	0.015 ± 0.004	0.013 ± 0.008	0.014 ± 0.009	0.014 ± 0.009
Morphometrics	0.302 ± 0.019	0.349 ± 0.162	0.302 ± 0.157	0.244 ± 0.109
Autoencoder	0.217 ± 0.010	0.220 ± 0.152	0.200 ± 0.117	0.193 ± 0.122
Pre-trained CLIP	0.319 ± 0.012	0.264 ± 0.057	0.337 ± 0.097	0.228 ± 0.096
GCD-CLIP	0.493 ± 0.078	0.583 ± 0.212	0.486 ± 0.157	0.294 ± 0.116
SimGCD-CLIP	0.520 ± 0.057	0.586 ± 0.060	<u>0.571 ± 0.129</u>	0.277 ± 0.137
GET	0.512 ± 0.035	0.610 ± 0.066	0.528 ± 0.136	0.313 ± 0.138
AnyMor-VLM (ours)	0.540 ± 0.048	0.582 ± 0.094	0.632 ± 0.146	<u>0.301 ± 0.169</u>

unsupervised, and novel-classes regimes. Specifically, the GCD family attains OA scores in the 61–63% range, exceeding pre-trained CLIP and urban morphometrics by 16–19% points. For novel classes, SimGCD-CLIP, GET, and GCD-CLIP secure either the best or the second-best results on ACC, ARI, and NMI. These results provide evidence that GCD techniques improve the recognition of established morphotypes and the discovery of held-out morphotypes under the evaluated benchmark protocol.

- (4) AnyMor-VLM delivers the best overall performance. The proposed AnyMor-VLM achieves the highest OA, ARI, and NMI on the urban-morphotype benchmark. Relative to the second-best GCD method, AnyMor-VLM improves OA by 2.10–2.53% points, ARI by 0.024, and NMI by 0.020, while also exhibiting lower standard deviations. Its most pronounced advantage lies in the unlabeled-class metrics, where it outperforms the runner-up by 3% points in accuracy, 0.069 in ARI, and 0.161 in NMI. It further ranks within the top three across all novel-class metrics. Collectively, these findings indicate that AnyMor-VLM comprehensively outperforms both urban-morphology and GCD baselines in the urban-morphotype benchmark.

To interrogate the contribution of each component, we conducted a series of ablation experiments over the three branches of AnyMor-VLM (Table 3). Within the task of open-world urban morphology classification, the visual-geometric branch emerges as the dominant contributor, achieving the highest OA, ARI, and NMI among the three single-branch variants. This is consistent with the underlying objective of the task, namely, to recognize the geometric regularities of urban form from figure-ground maps in order to differentiate urban morphotypes. The conceptual-semantic branch ranks second, corroborating the effectiveness of CLIP’s cross-modal alignment in coupling figure-ground representations with the lexical semantics of morphotype labels.

The map-description branch functions primarily as a source of complementary information. Taken alone, it exhibits limited discriminative power (OA = 29.81 ± 2.68%). We attribute it to two factors. 1) The generated descriptions tend to capture attributes that are too generic to be uniquely mapped onto specific morphotypes, and 2) a small fraction of descriptions contain factual inaccuracies (Figure S2). This low level of accuracy gives rise to a structural mismatch in the Hungarian clustering evaluation algorithm, thus resulting in its failure when calculating the accuracy for subsets. This indicates that map descriptions alone are insufficient to disambiguate the full set of urban morphotypes. Nevertheless, fusing this branch with the other two yields a substantial and stable improvement, reaching the highest OA of 64.87%, ARI of 0.465, and

Table 3. Ablation study of AnyMor-VLM with and without (w/o) each of its three branches, namely the conceptual-semantic (cs) branch, the visual-geometric (VG) branch, and the map-description (MD) branch. Means and standard deviations are computed over ten random 2:2:2 configurations. Bold values denote the best performance, and underlined values indicate the second-best.

Methods	OA%	ACC _L %	ACC _{UL} %	ACC _{novel} %
CS-branch	52.01 ± 5.52	68.88 ± 26.49	43.75 ± 15.35	40.29 ± 14.07
VG-branch	61.24 ± 5.65	68.19 ± 15.08	64.67 ± 22.10	49.07 ± 15.78
MD-branch	29.81 ± 2.68	–	–	–
w/o CS-branch	<u>64.02 ± 3.40</u>	73.06 ± 15.84	70.82 ± 24.42	46.71 ± 15.53
w/o VG-branch	57.89 ± 5.89	70.72 ± 11.60	58.35 ± 22.88	43.64 ± 15.63
w/o MD-branch	62.05 ± 3.45	69.57 ± 13.03	68.86 ± 22.03	46.69 ± 16.25
AnyMor-VLM	64.87 ± 3.62	74.05 ± 12.56	70.67 ± 21.75	48.35 ± 16.10
	ARI _{overall}	ARI _L	ARI _{UL}	ARI _{novel}
CS-branch	0.320 ± 0.041	0.619 ± 0.082	0.245 ± 0.062	0.228 ± 0.132
VG-branch	0.437 ± 0.065	0.610 ± 0.057	0.594 ± 0.154	0.276 ± 0.165
MD-branch	0.071 ± 0.019	0.220 ± 0.111	0.099 ± 0.122	0.205 ± 0.165
w/o CS-branch	<u>0.459 ± 0.055</u>	0.627 ± 0.119	0.675 ± 0.184	0.303 ± 0.174
w/o VG-branch	0.373 ± 0.066	0.562 ± 0.063	0.404 ± 0.262	0.247 ± 0.153
w/o MD-branch	0.445 ± 0.053	0.598 ± 0.065	0.625 ± 0.170	0.296 ± 0.190
AnyMor-VLM	0.465 ± 0.055	0.613 ± 0.101	0.663 ± 0.172	0.296 ± 0.191
	NMI _{overall}	NMI _L	NMI _{UL}	NMI _{novel}
CS-branch	0.389 ± 0.042	0.510 ± 0.092	0.306 ± 0.086	0.242 ± 0.118
VG-branch	0.520 ± 0.057	0.586 ± 0.060	0.571 ± 0.129	0.277 ± 0.137
MD-branch	0.137 ± 0.040	0.207 ± 0.109	0.106 ± 0.190	0.193 ± 0.143
w/o CS-branch	<u>0.535 ± 0.051</u>	0.603 ± 0.089	0.637 ± 0.165	0.303 ± 0.151
w/o VG-branch	0.440 ± 0.060	0.511 ± 0.051	0.389 ± 0.220	0.255 ± 0.118
w/o MD-branch	0.531 ± 0.051	0.589 ± 0.048	0.600 ± 0.155	0.300 ± 0.162
AnyMor-VLM	0.540 ± 0.048	0.582 ± 0.094	0.632 ± 0.146	0.301 ± 0.169

NMI of 0.540. Consequently, the utility of the map-description branch is realized in conjunction with the visual-geometric and conceptual-semantic branches.

4.3 Discovering known and unknown urban forms worldwide

We deployed the AnyMor-VLM to cities or regions in the world to search for unknown urban morphotypes. In this phase, the model was retrained with the full set of our benchmark data and attained an overall accuracy of 0.853 on the benchmark. We mapped the geographical distribution of urban forms across four regions: Barcelona in Spain, Singapore, Stratton Region in Australia, and Nairobi in Kenya.

In the low-dimensional embedding space, the semi-supervised K-Means++ algorithm identified nine clusters from the given regions and the benchmark dataset (Figure 5(a)), comprising six known morphotypes and three unknown morphotypes. The unknown type 1 exhibits high spatial density and mixed-shape building footprints (incorporating both large and small structures) (A1-3 in Figure 5(b)). By applying the attention mechanism of the VLM in the map-description branch, the most important map descriptions for the unknown type 1 are 'mix', 'density', 'organic' and 'planned' (A4 in Figure 5(b)). Thus, we reason that it is a kind of mixed-type dense urban form. The unknown type 2 also exhibits mixed-shape building footprints, albeit with moderate spatial density, as evidenced by a greater proportion of non-built areas (B1-3 in Figure 5(b)). The keywords of the corresponding map descriptions are 'mix', 'organ', 'dispersed', 'suburban', etc., which suggest that it is a suburban-like mixed and dispersed type (B4 in Figure 5(b)). The unknown type 3 corresponds to urban forms characterized by the lowest observed spatial density, as illustrated in C1-3 in Figure 5(b). These forms are distinguished by the presence of sparse buildings and roads. Their keywords identified by the VLM support the pattern, such as 'low', 'density', 'organic', and 'rural' (C4 in Figure 5(b)). Hence, we infer that the unknown type 3 are rural-like low-density organic urban forms.

As illustrated in Figure 6, the AnyMor-VLM model successfully identifies both known morphotypes and unknown morphotypes in Barcelona and Stratton. The geographic distribution of urban forms in Barcelona city reveals its spatial structure (Figure 6(a,b)). At the core of Barcelona's urban area lies the industrial city planning, a design that has gained global recognition (Figure 6(a) Zone Orange). The spatial pattern exhibited is of a regular and ordered nature (Figure 6(b)). In the northern region, the urban area extends to high-density and partially ordered pre-industrial city planning. In the southern suburban areas, numerous urban forms of unknown type 2 are distributed, corresponding to industrial land (numerous factories) or urban-suburban transition zones shown in the remote sensing satellite imagery. The selection of Stratton and its surrounding areas as the analysis region was due to the presence of the renowned urban form of garden city planning. As illustrated in Zone Blue of Figure 6, the model effectively identifies the urban form of garden city planning in this area. Furthermore, the model identified a substantial number of the unknown type 3 (Zone Green) and a comparatively minor number of the unknown type 1 (Zone Red) areas distributed across the region, corresponding to the rural and urban areas, respectively.

Furthermore, the three identified unknown urban morphotypes unveil the urban-rural transition in Singapore. Figure 7(a) indicates that Singapore is predominantly covered by the urban forms of the three unknown morphotypes. Subsequent analysis of their distribution in the low-dimensional representation space suggests that the topological proximity of these three unknown types indicates their similar morphological characteristics with gradual transitions along the dashed arrows (Figure 7(b)). The integration of geographic distribution with their examples (Figure 7(a,c)) enables the identification of the unknown type 1–3 as follows: high-density mixed-use urban areas, medium-density mixed-use urban areas (industrial zones or transition zones), and low-density non-urban areas (rural areas or natural land cover within urban areas). This reasoning is consistent with the geographical distribution of Singapore's urban-suburban transition. The analyses of urban morphology mapping in two entirely new regions demonstrate the model's generalizability, unaffected by geographical heterogeneity.

4.4 Applications

- Delineating morphological urban areas

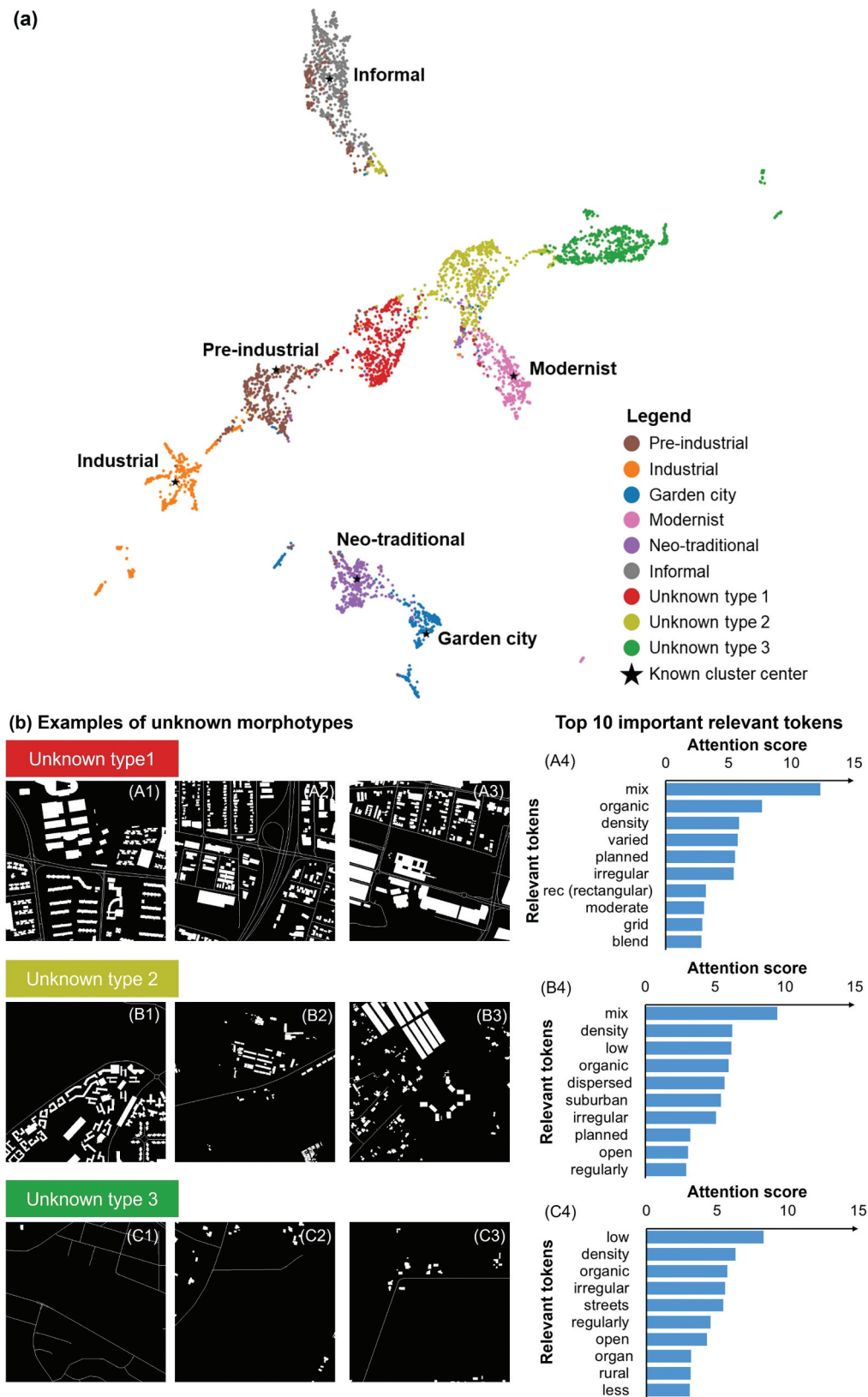


Figure 5. (a) Visualized clusters of urban morphological representations using UMAP dimensionality reduction technique. Six known morphotypes and three unknown morphotypes are shown in different colors. (b) Interpreting the three unknown urban morphotypes via calculating the token attention scores of map descriptions. A1-3, B1-3, C1-3 are corresponding examples of the three types. A4, B4, C4 show the top 10 important tokens related to spatial layout in the map descriptions. Some tokens are interpreted into words by querying the context of the corresponding map descriptions, such as 'rec' as 'rectangular'.

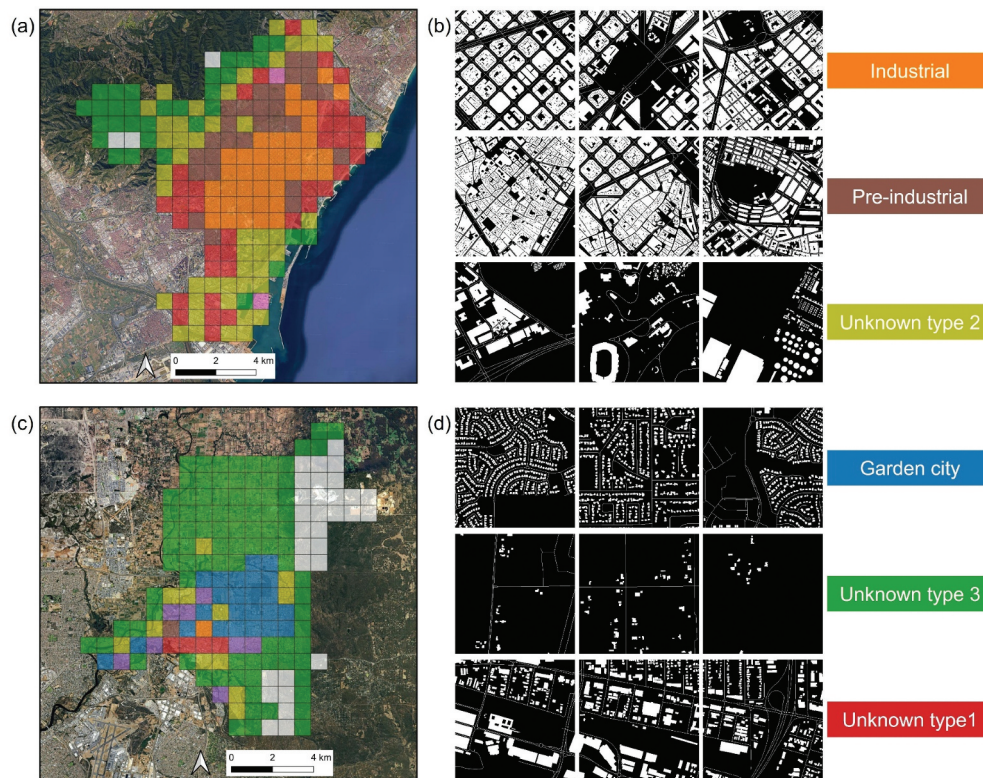


Figure 6. Geographic distribution of the identified urban morphotypes and representative examples in Barcelona, Spain (a-b) and Stratton and surrounding areas, Australia (c-d).

The arbitrary urban morphotypes revealed by the AnyMor-VLM model facilitate certain related applications, such as morphological urban area delineation. This application involves defining cities and forms the basis of many urban studies (Dong et al. 2024). Currently, the mainstream method is to use satellite imagery to delineate urban areas by identifying the morphology of human settlements (Taubenböck et al. 2019). This study excludes areas of rural forms and defines regions adjacent to the dense urban forms as morphological urban areas. Figure 7 shows the morphological urban areas in Singapore generated by the urban forms mapped in Figure 7(a). The resultant map is consistent with the reference layer delineating urban and suburban areas in Singapore (Blei and Angel 2021), which demonstrates the applicability of our model to this task. Furthermore, it enables the identification of internal urban spatial structures of cities (Anas, Arnott, and Small 1998). For example, Figure 6(a) illustrates Barcelona's classic urban structure centered around the Cerdà plan (Aibar and Bijker 1997; Santasusagna Riu et al. 2021). This provides a deeper understanding of a city's internal organization.

- Mapping and profiling slums

Another application of the identified urban morphotypes is urban slum mapping. This task is closely linked to the sustainable development goal 11 (Tu et al. 2024) and is a key focus of global urban research, especially regarding slums in Africa (Büttner et al. 2025; C. Li et al. 2025). Inferring urban slums from urban morphology is a common research approach (Kuffer et al. 2024). The proposed model successfully identified slums in Nairobi, Kenya. It is notable that data from this region was never included in the training dataset. The model distinguished between various forms of urban morphology in the area, including the slums with informal urban forms. The geographic distribution of urban morphology in Nairobi reveals four contiguous zones featuring informal settlement patterns (Figure 8(a) Zone Grey). Of particular interest are four notable slum areas, which were selected as exemplary cases across these contiguous zones, including Kawangware,

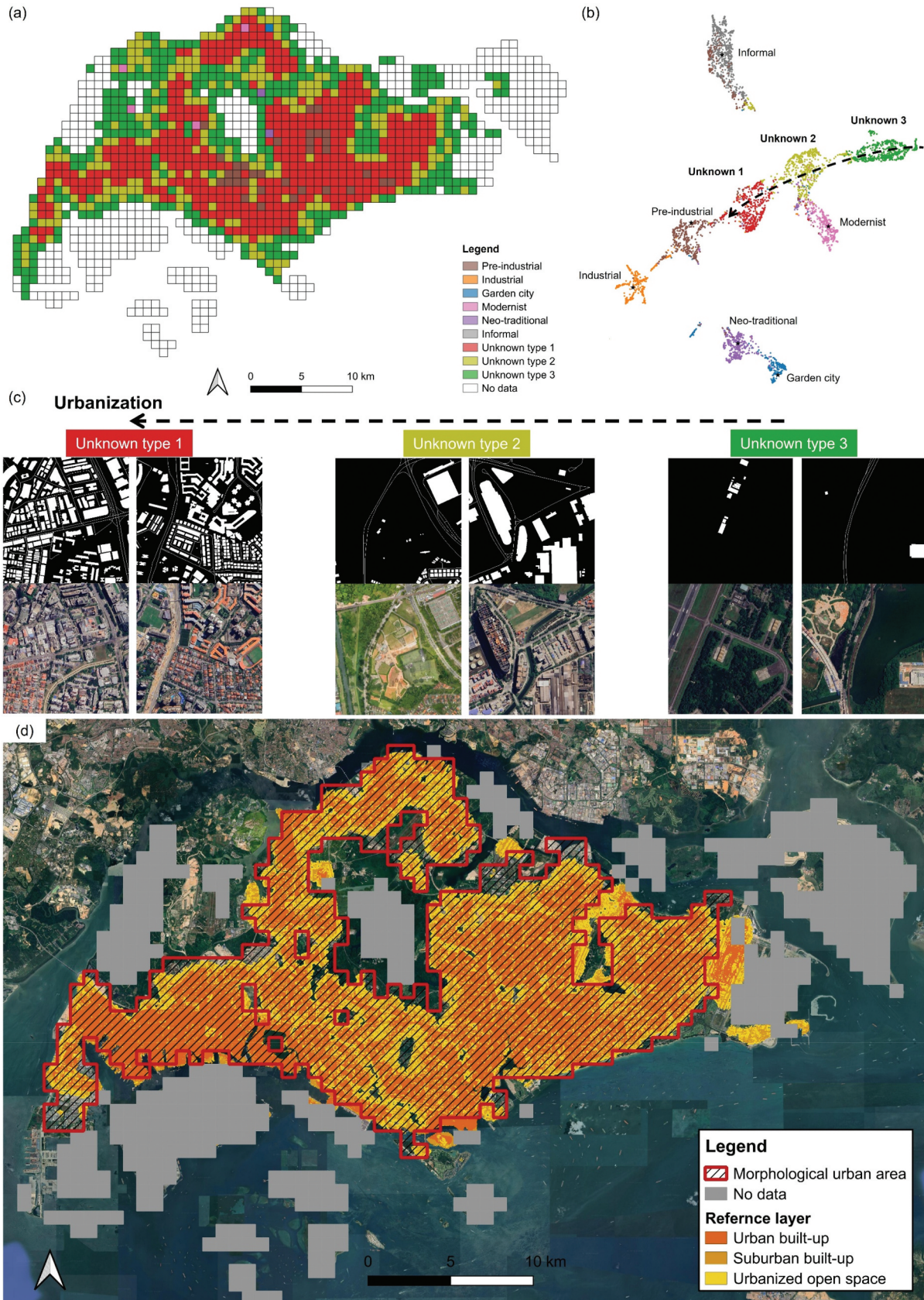


Figure 7. Urban-suburban transition revealed by the three unknown morphotypes in Singapore, whose data are never included in the training data. (a) Geographic distribution of urban morphology in Singapore, which shows the urban areas (red zone). (b) Urban-suburban transition projected in the low-dimensional space (the dotted line with arrows). (c) Examples of the three unknown morphotypes in Singapore, following the pattern of urban-suburban transition. (d) Delineating morphological urban areas using the urban morphology map generated by AnyMor-VLM. The reference layer of urban and suburban areas is collected from the Atlas of urban expansions (Blei and Angel 2021).

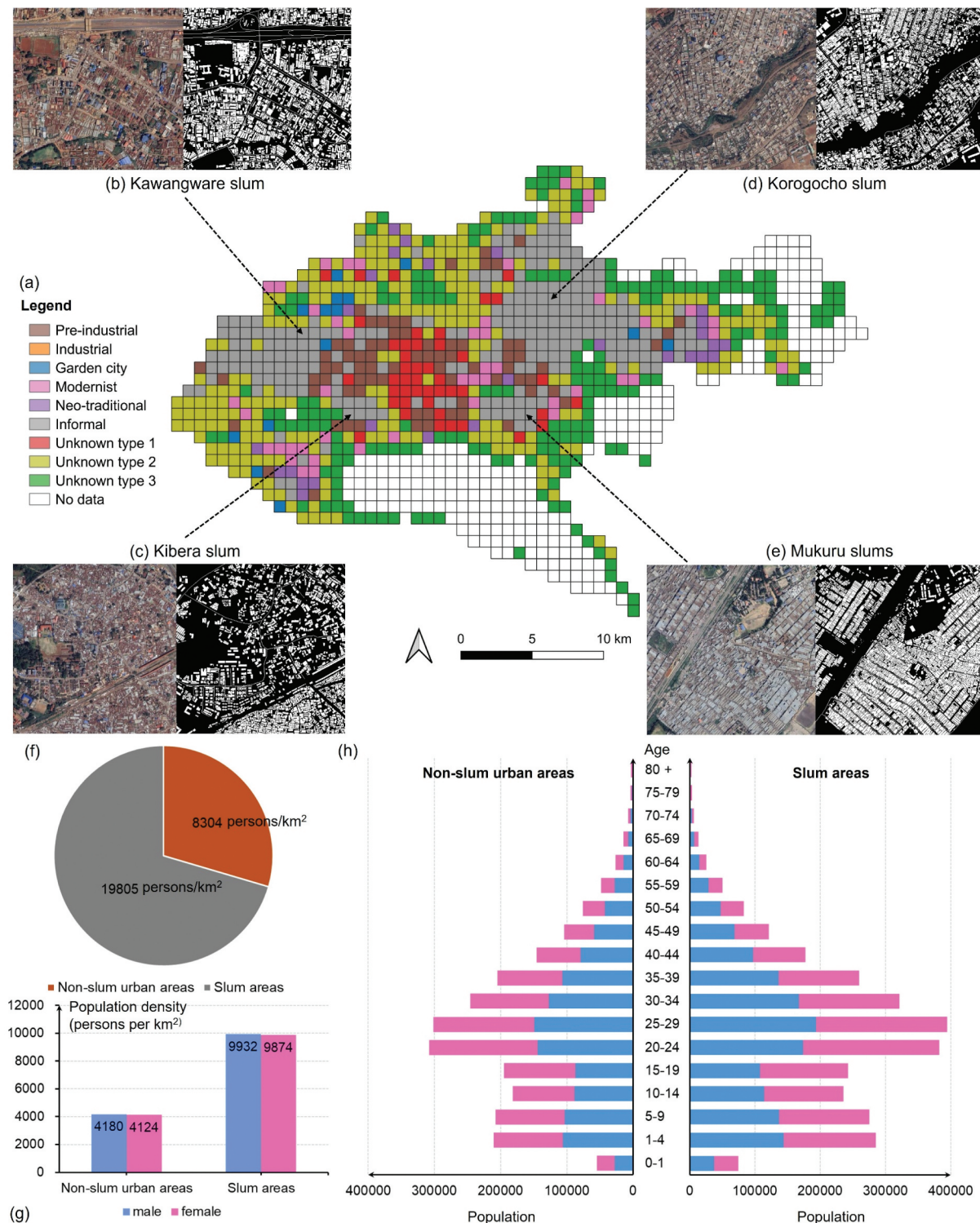


Figure 8. Mapping and profiling slums in Nairobi, Kenya using the urban morphology map generated by AnyMor-VLM. (a) Geographic distribution of urban morphology identified by AnyMor-VLM. Four examples are illustrated, including (b) Kawangware – a low-income residential area, (c) Kibera slum – Africa’s largest slum, (d) Korogocho – one of the largest slum neighbourhoods of Nairobi. (e) Mukuru – a collection of slums in the city. The demographics in the slum areas and non-slum urban areas are analysed by integrating the WorldPop population data with the urban morphology map. The difference between the slums and the non-slum urban areas is revealed in terms of (f) the overall population density, (g) the sex structure, as well as (h) the age structure.

Korogocho, Mukuru and the notable Kibera (Africa's largest slum) (Figure 8(b–e)). Consequently, the model demonstrates potential for identifying slums.

The demographics of the slums can be further analyzed by integrating the mapped slum areas with the WorldPop population data. The population density in slums is approximately double that of ordinary urban areas (Figure 8(f)), revealing the issue of high-density living conditions in these informal urban forms. This inequality is reflected rather in the age structure (Figure 8(g,h)). In comparison with non-slum urban areas, slums are home to a larger proportion of middle-aged and young people (under 45). This phenomenon may be attributed to the absence of upward social mobility within contemporary Nairobi society, a factor that hinders the mobility of current and future generations of workforce seeking to escape the slums.

5. Discussion

5.1 Benefits of open-world urban morphology classification

Urban morphology studies face the challenge of identifying unknown urban morphotypes, such as the existing but undefined urban forms across the globe, and the evolving future urban forms that are yet to be delineated. Whilst traditional urban morphotypes can be comprehensively covered through geometry-based typology systems, e.g. the gridiron, radial, organic system (W. Chen, Wu, and Biljecki 2021) or landscape-based typology systems, e.g. local climate zones (Zhu et al. 2022), these systems only coarsely characterize partial morphological features. They oversimplify the complexity of urban form and lack connections to existing urban morphology theories (Fleischmann et al. 2026). A significant number of urban morphotypes that are crudely defined as 'no pattern' yet actually contain deeper underlying patterns still exist globally. Furthermore, as complex adaptive systems, cities are subject to continuous evolution in response to societal demands, manifesting in their evolving urban forms. For instance, contemporary urban designers are actively pursuing the development of more efficient 15-minute compact city planning, sustainable urban forms, as well as smart cities integrating advanced technologies. A static typology system for urban morphology is ill-equipped to accommodate these emerging patterns. The central argument of this paper is that the introduction of a scalable, open-world typology system for urban morphology is imperative.

Our proposed open-world paradigm requires assessment models to demonstrate classification capabilities for known and novel urban morphotypes. Moreover, this work establishes a benchmark for urban morphotype under this paradigm, allowing future comparative studies to evaluate the capabilities of different models. It is posited that this challenging paradigm has the potential to improve the performance of the field of urban morphology classification models.

5.2 Integration of domain-knowledge awareness and data-driven capabilities

This paper argues that evaluating open-world urban forms requires an approach with both domain knowledge awareness and data-driven capabilities. Conventional supervised classification methods are developed within a framework driven by domain knowledge. However, this framework is susceptible to causing the classification methods to fall into the pitfalls of overfitting and local optima. The prevailing unsupervised clustering approaches, on the other hand, adopt a data-driven perspective, focusing on the intrinsic characteristics of geospatial data itself. However, geospatial data inherently contains high-dimensional information, which is mostly irrelevant to urban morphology. For instance, the extracted characteristics from figure-ground maps are frequently interpreted as the quantity and distribution of black-and-white pixels within an image (Moosavi 2022). Consequently, the data-driven approach needs to integrate domain knowledge to prioritize information pertinent to urban morphology.

The domain-knowledge-aware data-driven idea is manifested by several aspects of the proposed AnyMor-VLM model. The design of our model emphasizes the characteristics of the data itself while concomitantly integrating domain knowledge. The AnyMor-VLM principally accomplishes the data-driven urban morphological feature extraction through unsupervised learning techniques, including the clustering and self-supervised contrastive learning objectives. Concurrently, it projects the high-dimensional data features onto an urban morphological space of the domain knowledge via supervised learning objective and natural

language-guided representation learning. Furthermore, the multi-objective optimization is applied to integrate the two parts and plays a role of balancing the domain knowledge learning and data-feature learning.

Specifically, within the multi-objective optimization, our domain-knowledge-aware data-driven approach can be conceptualized as a weighted average of supervised and unsupervised learning. The determination of this weighted average is influenced by a hyperparameter λ . Achieving a model capable of recognizing known types of domain knowledge and discovering novel types, a nuanced balance is required between the relative weights of supervised and unsupervised learning. Consequently, the parameter sensitivity of λ was investigated during the model development phase. As illustrated in Figure 9, at a value of λ equal to 0.15 or 0.2, the overall accuracy achieves a peak. Moreover, when $\lambda = 0.15$, the three accuracies for subsets are found to be relatively balanced. Consequently, an unsupervised-dominant, supervised-supplementary configuration is observed to yield optimal results, thereby facilitating a domain-knowledge-aware data-driven framework.

5.3 Pros and cons of the map-description branch

The map-description branch provides a modest but consistent improvement. Across the 10 ablation experiments, removing this branch reduces OA from $64.87\% \pm 3.62\%$ to $62.05\% \pm 3.45\%$, an absolute decrease of 2.82% points (Table 3). Improvements are observed for all class groups, reaching 4.48, 1.81, and 1.66% points for the labeled, unlabeled, and novel classes, respectively. These results indicate that textual descriptions provide complementary morphological cues, although the visual-geometric branch remains the primary source of classification and category-discovery performance.

The branch introduces limited trainable complexity because GPT-4o is used only during preprocessing. Descriptions are generated once, cached, and encoded through the existing CLIP text encoder; the principal architectural addition is therefore the projection of the description representation within the fusion module. Nevertheless, the complete pipeline is not cost-free. Caption generation for new samples depends on an external multimodal model, entails computational or API costs, and may occasionally produce generic or inaccurate descriptions. The map-description branch is lightweight in terms of trainable architecture.

The branch also provides a form of model explainability that is particularly relevant to urban morphology. As shown in Figure 5(b), attributes such as density, regularity, and spatial arrangement correspond to concepts commonly used by urban morphologists to describe and compare urban forms. The generated descriptions therefore provide a human-readable account of the morphological evidence supplied to the model. They can be inspected alongside predictions to infer the morphological meanings of novel morphotypes identified by AnyMor-VLM. This interpretability is complementary to the numerical embeddings of the visual-geometric and conceptual-semantic branches. However, the branch does not make the complete model fully interpretable. Because the three representations are concatenated and transformed through

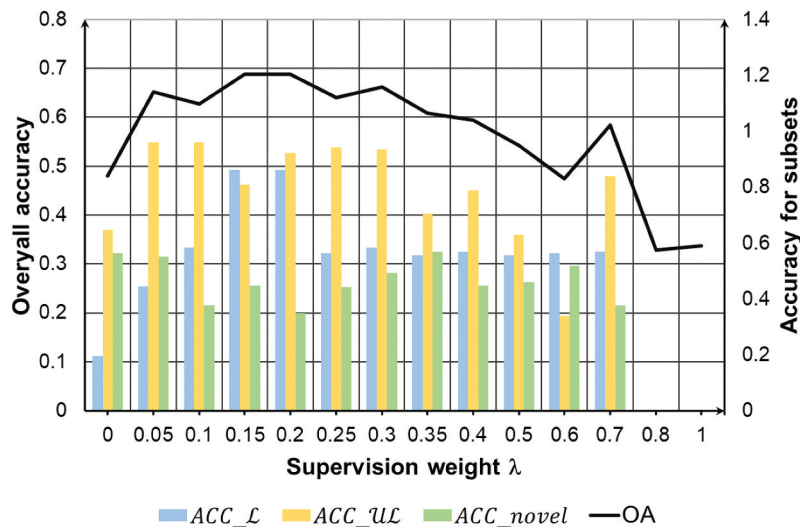


Figure 9. Parameter analysis for the weight of supervised learning and unsupervised learning λ .

a nonlinear fully connected projection, the raw fusion-layer parameters cannot be interpreted as scalar importance weights for individual branches. Similarly, although the CLIP text encoder uses an attention-based transformer architecture, attention values alone should not be treated as definitive explanations of the final prediction. A more rigorous analysis of token- and branch-level contributions remains an important direction for future work.

5.4 Limitations and future opportunities

Representational scope. The findings should be interpreted within the scope of binary figure-ground maps. In the Conzenian tradition, urban form is historically constituted through the interaction of town plans, building fabrics, land and building utilization, and successive morphological periods (Conzen 1960; Moudon 1997). Figure-ground maps capture only the two-dimensional relationship between built mass and open space. They make geometric properties such as density, grain, spacing, orientation, and compactness visible, but do not reveal land use, tenure, street hierarchy, infrastructure, historical sequence, or the social processes that produced these patterns. Visually similar configurations may therefore arise from different historical and institutional trajectories. The categories identified here are geometry-based morphotypes, but not comprehensive historical-genetic or socio-functional urban types. This restricted representation nevertheless provides a standardized input that can be generated consistently across cities and enables a focused test of whether built-open-space geometry alone supports the recognition and discovery of morphotypes (J. Wang, Huang, and Biljecki 2024). Excluding place names, coordinates, land-use variables, and other contextual information also helps limit location-specific and functional shortcuts. A fuller morphological interpretation would require additional evidence, including street and plot evolution, land-use change, building replacement cycles, planning histories, and locally grounded knowledge.

Performance. Although AnyMor-VLM outperforms the strongest baseline by 2.10–2.53% points in OA, 0.024 in ARI, and 0.020 in NMI, its OA of $64.87\% \pm 3.62\%$ remains below the longer-term objective of reliable global urban-morphology classification. The failure cases reported in Supplementary Section S4 and Figure S5 reflect both overlap among geometry-based morphotypes and information loss when complex spatial configurations are summarized as short descriptions. The current implementation is a proof of concept for applying generalized category discovery to urban morphology. We encourage more researchers to explore advanced open-world urban morphology classification methods. Further improvements should focus on enhancing the classification capabilities for known categories as well as the generalizability for unknown categories. For example, more advanced, pre-trained, large, multi-modal models that can simultaneously accommodate multiple data modalities, such as text, images, audio, and video, offer greater potential for advancement due to their stronger general classification capabilities.

Taxonomy and scope of contribution. The six categories constitute an operational benchmark taxonomy, not a universal typology. Several are rooted in planning and design traditions primarily codified in European and North American contexts and may not adequately represent locally specific, hybrid, or transitional forms elsewhere. Extending the framework will require culturally broader and hierarchically structured taxonomies, geographically balanced validation, community-contributed annotations, and collaboration with scholars familiar with regional planning and morphological traditions.

Spatial units. Fixed 400×400 m grids provide consistent image dimensions and facilitate controlled comparison using pretrained vision-language models. However, morphological legibility depends on the size and shape of the analytical unit, and individual cells may contain mixed or transitional forms. The assigned label should be interpreted as the dominant visible morphology. J. Wang, Huang, and Biljecki (2024) similarly show that no grid size is universally optimal, while recent work on hierarchical morphotopes emphasizes spatial continuity, nesting, and scale-dependent morphological character (Fleischmann et al. 2026). Future studies should examine adaptive or morphology-defined boundaries, hierarchical and multi-scale representations, and soft or multilabel classification. Incorporating building height and other three-dimensional information would further distinguish urban fabrics that appear similar in plan but differ substantially in vertical structure.

External geographic information. The present study intentionally excludes geographic coordinates, city names, case identifiers, and location-aware foundation models because its objective is to learn from

morphological geometry. Future work may incorporate contextual variables such as points of interest or land use, but such extensions should carefully control location-label correlations introduced by geographically clustered sampling and labeling procedures.

Trained GPT-4o. To reduce explicit information leakage from GPT-4o, inputs were anonymized, prompts were restricted to visually observable geometry, and the generated descriptions were reviewed for unsupported geographic references. No city names, coordinates, or country identifiers were identified, and all descriptions are provided in the data repository. Nevertheless, because GPT-4o's pretraining corpus is not publicly auditable, implicit recognition or memorization of characteristic urban patterns cannot be completely excluded. Future evaluations should compare model-generated descriptions with independently produced expert descriptions and with open or domain-specific captioning models.

Potential geographic biases. The results are also conditioned by the geographic coverage of the benchmark, the uneven completeness of OpenStreetMap, and biases inherited from Internet-scale pretrained models. OSM provides an openly licensed and comparatively consistent source of building footprints, but its completeness and positional accuracy vary across regions, particularly in rapidly changing and resource-constrained areas (Herfort et al. 2023). Although participatory mapping can achieve high completeness in some informal settlements (Yeboah et al. 2021), observed regional differences may still partly reflect mapping practices rather than urban morphology alone. Future pipelines should combine OSM with satellite-derived building footprints and explicitly propagate data-quality uncertainty. CLIP and GPT-4o may also reflect uneven geographic, linguistic, and cultural representation in their training corpora. Anonymizing the inputs reduces explicit geographic dependence but cannot eliminate biases embedded in pretrained representations. Consistent with Drici and Carpio-Pinedo (2025), future work should prioritize geographically stratified evaluation, locally grounded morphotype development, improved representation of informal and data-poor urban conditions, and sustained participation by regional experts.

6. Conclusion

This study first introduces an evaluation paradigm for open-world urban morphology classification and a benchmark dataset to assess the model performance on both known and unknown urban morphotypes. Under this paradigm, overall accuracy is divided into three distinct measures, i.e. labeled-class accuracy, unlabeled-class accuracy, and novel-class accuracy, to evaluate model performance across different classes according to label availability.

Driven by this challenging evaluation paradigm, this paper proposes a novel triple-branch framework named AnyMor-VLM for open-world urban morphology classification. The innovative framework integrates supervised learning guided by natural language and unsupervised learning, thereby endowing the model with both domain-knowledge awareness and data-driven capabilities. The model incorporates three branches, each characterized by a distinct progressive fine-tuning strategy. These strategies facilitate the acquisition of domain-specific knowledge while preserving the model's existing generalizability.

Comprehensive experiments were conducted, showing the superior performance of the proposed AnyMor-VLM in comparison to existing state-of-the-art methods on the urban morphotype benchmark dataset. It exhibits strong generalizability, enabling its extension to untrained regions for discovery of novel categories and analysis of urban form across different geographical contexts. And the two proof-of-concept applications, urban area delineation and slum mapping, demonstrate the effectiveness of the mapped urban morphotypes.

This work demonstrates the feasibility of bridging the current gap between existing artificial general intelligence technologies and complex domain knowledge. On the one hand, it proactively introduces new AI techniques into the realm of classic urban morphology research, thereby stimulating the discovery of novel domain knowledge. Conversely, it propels global urban studies in close conjunction with urban morphology, thus facilitating further exploration and advancing the realization of urban design for sustainable development.

Author contributions

CRediT: **Dongsheng Chen:** Conceptualization, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing; **Yao Yao:** Methodology, Writing – review & editing; **Xun Li:** Formal analysis, Writing – original draft; **Jinzhou Cao:** Methodology, Writing – original draft; **Yu Gu:** Formal analysis, Writing – review & editing; **Liqu Meng:** Formal analysis, Supervision, Writing – original draft.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability statement

The urban morphotype benchmark dataset and the generated map descriptions analysed in this paper are publicly available on Zenodo (<https://doi.org/10.5281/zenodo.21679157>).

Code availability

The code of AnyMor-VLM is available on GitHub for contiguous development (<https://github.com/DongshengChen9/AnyMor-VLM>).

References

- Aibar, E., and W. E. Bijker. 1997. "Constructing a City: The Cerdà Plan for the Extension of Barcelona." *Science, Technology, and Human Values* 22 (1): 3–30. <https://doi.org/10.1177/016224399702200101>.
- Anas, A., R. Arnott, and K. A. Small. 1998. "Urban Spatial Structure." *Journal of Economic Literature* 36:3 1426–1464.
- Arribas-Bel, D., and M. Fleischmann. 2022. "Spatial Signatures - Understanding (Urban) Spaces Through Form and Function." *Habitat International* 128:102641. <https://doi.org/10.1016/j.habitatint.2022.102641>
- Assran, M., M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat, and N. Ballas. 2022. "Masked Siamese Networks for Label-Efficient Learning." In *Computer Vision – ECCV 2022, Lecture Notes in Computer Science*, edited by S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, 456–473. Cham: Springer Nature Switzerland.
- Ateş, A. H., and U. Yurtsever. 2026. "An Overview Of Multimodal Learning: Concepts, Challenges, Applications And Datasets." *Konya Journal of Engineering Sciences* 14 (2): 917–950. <https://doi.org/10.36306/konjes.1716813>.
- Batty, M. 2024. *The Computable City: Histories, Technologies, Stories, Predictions*. Cambridge, Massachusetts: The MIT Press.
- Becht, E., L. McInnes, J. Healy, C.-A. Dutertre, Immanuel W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. 2019. "Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP." *Nature Biotechnology* 37 (1): 38–44. <https://doi.org/10.1038/nbt.4314>.
- Biljecki, F., and Y. S. Chow. 2022. "Global Building Morphology Indicators." *Computers, Environment and Urban Systems* 95:101809. <https://doi.org/10.1016/j.compenvurbsys.2022.101809>.
- Blei, A., and S. Angel. 2021. "Global monitoring with the Atlas of Urban Expansion." In X. Yang (Ed.), *Urban remote sensing*, 247–282. New Jersey, U.S.: John Wiley & Sons, Ltd.
- Boeing, G. 2018. "Measuring the Complexity of Urban Form and Design." *Urban Design International* 23 (4): 281–292. <https://doi.org/10.1057/s41289-018-0072-1>.
- Bruno, M., H. P. Monteiro Melo, B. Campanelli, and V. Loreto. 2024. "A Universal Framework for Inclusive 15-Minute Cities." *Nature Cities* 1 (10): 633–641. <https://doi.org/10.1038/s44284-024-00119-4>.
- Büttner, N., S. Stalder, M. Volpi, E. Suel, and K. Harttgen. 2025. "Large-Scale Slum Mapping in Sub-Saharan Africa's Major Cities: Remote Sensing and Deep Learning Reveal Strong Slum Growth in the Urban Periphery Between 2016 and 2022." *Habitat International* 161:103403. <https://doi.org/10.1016/j.habitatint.2025.103403>.
- Cao, J., J. Chen, X. Wang, W. Huang, D. Chen, T. Zhao, W. Tu, and Q. Li. 2026. "UrbanMMCL: Urban Region Representations via Multi-Modal and Multi-Graph Self-Supervised Contrastive Learning." *ISPRS Journal of Photogrammetry & Remote Sensing* 232:75–93. <https://doi.org/10.1016/j.isprsjprs.2025.11.012>.
- Cao, K., M. Brbic, and J. Leskovec. 2022. "Open-World Semi-Supervised Learning."
- Chacón, J. E., and A. I. Rastrojo. 2023. "Minimum Adjusted Rand Index for Two Clusterings of a Given Size." *Advances in Data Analysis and Classification* 17 (1): 125–133. <https://doi.org/10.1007/s11634-022-00491-w>.
- Chen, D., Y. Feng, X. Li, M. Qu, P. Luo, and L. Meng. 2025. "Interpreting Core Forms of Urban Morphology Linked to Urban Functions with Explainable Graph Neural Network." *Computers, Environment and Urban Systems* 118:102267. <https://doi.org/10.1016/j.compenvurbsys.2025.102267>.

- Chen, W., H. Huang, S. Liao, F. Gao, and F. Biljecki. 2024. "Global Urban Road Network Patterns: Unveiling Multiscale Planning Paradigms of 144 Cities with a Novel Deep Learning Approach." *Landscape and Urban Planning* 241:104901. <https://doi.org/10.1016/j.landurbplan.2023.104901>.
- Chen, W., A. N. Wu, and F. Biljecki. 2021. "Classification of Urban Morphology with Deep Learning: Application on Urban Vitality." *Computers, Environment and Urban Systems* 90:101706. <https://doi.org/10.1016/j.compenvurbsys.2021.101706>.
- Chen, Z., and B. Liu. 2018. "Open-World Learning." In *Lifelong Machine Learning*, edited by Z. Chen and B. Liu, 77–89. Cham: Springer International Publishing.
- Congress for the New Urbanism. 2000. "Charter of the New Urbanism." *Bulletin of Science, Technology & Society* 20 (4): 339–341. <https://doi.org/10.1177/027046760002000417>.
- Conzen, M. R. G. 1960. "Alnwick, Northumberland: A Study in Town-Plan Analysis." *Transactions and Papers (Institute of British Geographers)* (27): iii–122. <https://doi.org/10.2307/621094>.
- Conzen, M. R. G. 2004. *Thinking About Urban Form: Papers on Urban Morphology, 1932–1998*. Oxford / Bern / New York: Peter Lang.
- D'acci, L. S., D. Banister, and R. W. White. 2024. "Liveable Urban Forms: Planning, Self-Organisation, and a Third Way (Isobenefit Urbanism)." *Humanities and Social Sciences Communications* 11 (1): 578. <https://doi.org/10.1057/s41599-024-02975-w>.
- Debray, H., M. Gassilloud, R. Lemoine-Rodríguez, M. Wurm, X. Zhu, and H. Taubenböck. 2025. "Universal Patterns of Intra-Urban Morphology: Defining a Global Typology of the Urban Fabric Using Unsupervised Clustering." *International Journal of Applied Earth Observation and Geoinformation* 141:104610. <https://doi.org/10.1016/j.jag.2025.104610>.
- Dong, L., F. Duarte, G. Duranton, P. Santi, M. Barthelemy, M. Batty, L. Bettencourt, et al. 2024. "Defining a City — Delineating Urban Areas Using Cell-Phone Data." *Nature Cities* 1 (2): 117–125. <https://doi.org/10.1038/s44284-023-00019-z>.
- Drici, H., and J. Carpio-Pinedo. 2025. "Urban Land Use Mix and AI: A Systematic Review." *Cities* 165:106102. <https://doi.org/10.1016/j.cities.2025.106102>.
- Fleischmann, M., A. Feliciotti, and W. Kerr. 2022. "Evolution of Urban Patterns: Urban Morphology as an Open Reproducible Data Science." *Geographical Analysis* 54 (3): 536–558. <https://doi.org/10.1111/gean.12302>.
- Fleischmann, M., A. Feliciotti, O. Romice, and S. Porta. 2022. "Methodological Foundation of a Numerical Taxonomy of Urban Form." *Environment & Planning B: Urban Analytics & City Science* 49 (4): 1283–1299. <https://doi.org/10.1177/23998083211059835>.
- Fleischmann, M., K. Samardzhiev, A. Brázdová, D. Dančejová, and L. Winkler. 2026. "The Hierarchical Morphotype Classification: A Theory-Driven Framework for Large-Scale Analysis of Built Form." *Cities* 174:107047. <https://doi.org/10.1016/j.cities.2026.107047>.
- He, Z., Y. Liu, and K. Han. 2025. "Category Discovery: An Open-World Perspective."
- Herfort, B., S. Lautenbach, J. Porto de Albuquerque, J. Anderson, and A. Zipf. 2023. "A Spatio-Temporal Analysis Investigating Completeness and Inequalities of Global Urban Building Data in OpenStreetMap." *Nature Communications* 14 (1): 3985. <https://doi.org/10.1038/s41467-023-39698-6>.
- Howard, E. 1902. *Garden Cities of To-Morrow*. London: Routledge.
- Huang, J., X. X. Lu, and J. M. Sellers. 2007. "A Global Comparative Analysis of Urban Form: Applying Spatial Metrics and Remote Sensing." *Landscape and Urban Planning* 82 (4): 184–197. <https://doi.org/10.1016/j.landurbplan.2007.02.010>.
- Kuffer, M., A. Abascal, R. Engstrom, D. R. Thomson, G. Tregonning, A. Shonowo, Q. Zhao, et al. 2024. "IDEAMAPS: Modelling Sub-Domains of Deprivation with EO and AI." IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium. Presented at the IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium Athens, Greece, 1562–1566.
- Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*. Vancouver, Canada: Curran Associates Inc, 18661–18673.
- Li, C., L. Yu, R. Ndugwa, A. J. Bailey, X. Zhang, F. Oloo, E. G. Chimimba, X. Deng, and J. Wright. 2025. "Mapping Urban Slums and Their Inequality in Sub-Saharan Africa." *Nature Cities* 2 (11): 1037–1048. <https://doi.org/10.1038/s44284-025-00276-0>.
- Li, J., C. Li, S. J. Quan, and Y. Liang. 2026. "Deciphering Urban Land Use: A Typological Analysis of Built Form Across 105 Urban Areas in the U.S. Using Representation Learning." *Land Use Policy* 161:107878. <https://doi.org/10.1016/j.landusepol.2025.107878>.
- Liu, X., F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang. 2023. "Self-Supervised Learning: Generative or Contrastive." *IEEE Transactions on Knowledge and Data Engineering* 35:1 857–876 doi:10.1109/TKDE.2021.3090866.
- Liu, Y., Z. He, and K. Han. 2025. "Hyperbolic Category Discovery." 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Nashville, TN), 9891–9900.
- Moosavi, V. 2022. "Urban morphology meets deep learning." In *Machine learning and the city*. John Wiley & Sons, Ltd. 379–392. <https://doi.org/10.1002/9781119815075.ch29>.
- Moudon, A. V. 1997. "Urban Morphology as an Emerging Interdisciplinary Field." *Urban Morphology* 1 (1): 3–10. <https://doi.org/10.5134/jum.v1i1.4047>.

- Oliveira, V., and S. Porta. 2025. "Quantitative and Qualitative Analysis in Urban Morphology: Systematic Legacy and Latest Developments." *Proceedings of the Institution of Civil Engineers - Urban Design and Planning* 178 (2): 75–87. <https://doi.org/10.1680/jurdp.24.00047>.
- Osco, L. P., Q. Wu, E. L. de Lemos, W. N. Gonçalves, Ana Paula Marques. Ramos, J. Li, and J. Marcato. 2023. "The Segment Anything Model (SAM) for Remote Sensing Applications: From Zero to One Shot." *International Journal of Applied Earth Observation and Geoinformation* 124:103540. <https://doi.org/10.1016/j.jag.2023.103540>.
- Ouldoughi, R., C.-W. Kuo, and Z. Kira. 2023. "CLIP-GCD: Simple Language Guided Generalized Category Discovery." *ACM Computing Surveys* 55 (10): 1–37. <https://doi.org/10.1145/3561381>.
- Parmar, J., S. Chouhan, V. Raychoudhury, and S. Rathore. 2023. "Open-World Machine Learning: Applications, Challenges, and Opportunities." *ACM Computing Surveys* 55 (10): 1–37. <https://doi.org/10.1145/3561381>.
- Santasusagna Riu, A., J. Tort Donada, M. T. Vadrí Fortuny, and V. Paül Carril. 2021. "Estimating Public Space Metrics from Nineteenth-Century Urban Cartography: Barcelona's Cerdà Plan of Urban Expansion." *Environment & Planning B: Urban Analytics & City Science* 48 (9): 2640–2655. <https://doi.org/10.1177/2399808320987562>.
- Stefanini, M., M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara. 2023. "From Show to Tell: A Survey on Deep Learning-Based Image Captioning." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 45 (1): 539–559. <https://doi.org/10.1109/TPAMI.2022.3148210>.
- Taubenböck, H., M. Weigand, T. Esch, J. Staab, M. Wurm, J. Mast, and S. Dech. 2019. "A New Ranking of the World's Largest cities—Do Administrative Units Obscure Morphological Realities?" *Remote Sensing of Environment* 232:111353. <https://doi.org/10.1016/j.rse.2019.111353>.
- Troisemaine, C., V. Lemaire, S. Gosselin, A. Reiffers-Masson, J. Flocon-Cholet, and S. Vaton. 2023. "Novel Class Discovery: An Introduction and Key Concepts."
- Tu, W., D. Chen, R. Cao, J. Xia, Y. Zhang, and Q. Li. 2024. "Towards SDG 11: Large-Scale Geographic and Demographic Characterisation of Informal Settlements Fusing Remote Sensing, POI, and Open Geo-Data." *ISPRS Journal of Photogrammetry & Remote Sensing* 217:199–215. <https://doi.org/10.1016/j.isprsjprs.2024.08.014>.
- Vaze, S., K. Han, A. Vedaldi, and A. Zisserman. 2022. "Generalized Category Discovery."
- Wang, E., Z. Peng, Z. Xie, F. Yang, X. Liu, and M.-M. Cheng. 2025. "GET: Unlocking the Multi-Modal Potential of CLIP for Generalized Category Discovery." *Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20296–20306.
- Wang, J., W. Huang, and F. Biljecki. 2024. "Learning Visual Features from Figure-Ground Maps for Urban Morphology Discovery." *Computers, Environment and Urban Systems* 109:102076. <https://doi.org/10.1016/j.compenvurbsys.2024.102076>.
- Wang, J., W. Xuan, H. Qi, Z. Chen, H. Chen, Z. Zheng, J. Xia, Y. Zhong, and N. Yokoya. 2026. "CityVLM: Towards Sustainable Urban Development via Multi-View Coordinated Vision–Language Model." *ISPRS Journal of Photogrammetry & Remote Sensing* 232:62–74. <https://doi.org/10.1016/j.isprsjprs.2025.11.030>.
- Wen, X., B. Zhao, and X. Qi. 2023. "Parametric Classification for Generalized Category Discovery: A Baseline Study." *Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision Paris, France*, 16590–16600.
- Whitehand, J. W. R. 2001. "British Urban Morphology: The Conzenian Tradition." *Urban Morphology* 5 (2): 103–109. <https://doi.org/10.51347/jum.v5i2.3896>.
- Yeboah, G., J. Porto de Albuquerque, R. Troilo, G. Tregonning, S. Perera, Syed A. K. Shifat. Ahmed, M. Ajisola, et al. 2021. "Analysis of OpenStreetMap Data Quality at Different Stages of a Participatory Mapping Process: Evidence from Slums in Africa and Asia." *ISPRS International Journal of Geo-Information* 10 (4): 265. <https://doi.org/10.3390/ijgi10040265>.
- Zhang, P., D. Ghosh, and S. Park. 2023. "Spatial Measures and Methods in Sustainable Urban Morphology: A Systematic Review." *Landscape and Urban Planning* 237:104776. <https://doi.org/10.1016/j.landurbplan.2023.104776>.
- Zhu, X. X., C. Qiu, J. Hu, Y. Shi, Y. Wang, M. Schmitt, and H. Taubenböck. 2022. "The Urban Morphology on Our Planet – Global Perspectives from Space." *Remote Sensing of Environment* 269:112794. <https://doi.org/10.1016/j.rse.2021.112794>.