UV-SAM: Adapting Segment Anything Model for Urban Village Identification

Xin Zhang[1], **Yu Liu**[2*], **Yuming Lin**[2], **Qingming Liao**[1], **Yong Li**[2]

[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]Department of Electronic Engineering, Tsinghua University, Beijing, China
zhangxin4087@163.com, liuyu2419@126.com

Abstract

Urban villages, defined as informal residential areas in or around urban centers, are characterized by inadequate infrastructures and poor living conditions, closely related to the Sustainable Development Goals (SDGs) on poverty, adequate housing, and sustainable cities. Traditionally, governments heavily depend on field survey methods to monitor the urban villages, which however are time-consuming, labor-intensive, and possibly delayed. Thanks to widely available and timely updated satellite images, recent studies develop computer vision techniques to detect urban villages efficiently. However, existing studies either focus on simple urban village image classification or fail to provide accurate boundary information. To accurately identify urban village boundaries from satellite images, we harness the power of the vision foundation model and adapt the Segment Anything Model (SAM) to urban village segmentation, named UV-SAM. Specifically, UV-SAM first leverages a small-sized semantic segmentation model to produce mixed prompts for urban villages, including mask, bounding box, and image representations, which are then fed into SAM for fine-grained boundary identification. Extensive experimental results on two datasets in China demonstrate that UV-SAM outperforms existing baselines, and identification results over multiple years show that both the number and area of urban villages are decreasing over time, providing deeper insights into the development trends of urban villages and sheds light on the vision foundation models for sustainable cities. The dataset and codes of this study are available at https://github.com/tsinghua-fib-lab/UV-SAM.

Introduction

As a representative type of informal settlement, urban villages are densely populated neighborhoods in both the outskirts and the downtown segments of major Chinese cities, typically consisting of older low-rise buildings and narrow alleyways (Wikipedia contributors 2023), as shown in Figure 1. On the one hand, urban villages provide affordable housing options for migrant workers and low-income citizens, contributing to the socioeconomic fabric of cities. On the other hand, urban villages often face challenges related to inadequate infrastructure, limited access to public services, and poor living conditions (Chen et al. 2021). Hence, aligning well with United Nations' 11th Sustainable Development Goal (SDG 11)"Making cities and human settlements inclusive, safe, resilient and sustainable" (Nath 2016), accurately identifying urban villages is essential for both urban planning and governance in future sustainable cities.

(a) Narrow alleyways (b) Older low-rise buildings

Figure 1: Examples of urban villages identified from satellite images, with appearance characteristics provided. The red part represents the urban village areas.

Traditionally, urban village identification heavily depends on field surveys and manual mapping (Zheng et al. 2009), where urban planners would visit different areas, collect socioeconomic data, and visually identify urban village boundaries. While such methods provide valuable insights, they are time-consuming, labor-intensive, and limited in spatiotemporal coverage. In recent years, exploring computer vision techniques with satellite images for urban villages has gained significant attention. Most studies build image classification models to classify whether a given satellite image contains an urban village (Chen et al. 2022; Fan et al. 2022a,b; Xiao et al. 2023) without boundaries identified, while others explore semantic segmentation models to identify urban village boundaries in satellite images (Mast, Wei, and Wurm 2020; Pan et al. 2020; Chen et al. 2019). However, due to the complex background interference in satellite images and the lack of well-defined boundaries between urban villages and surrounding neighborhoods, existing studies perform poorly in providing accurate urban village boundaries, which further hinders the estimation of the areas and expansions of urban villages (Kirillov et al. 2023).

*Corresponding author

Moreover, limited labeled data in urban villages also make segmentation models prone to overfit and fail to generalize to noisy satellite images, e.g., occlusion and seasons.

Meanwhile, owing to training on over one billion images, the recent vision foundation model of the Segment Anything Model (SAM) exhibits remarkable generalization capabilities as well as high mask quality for category-agnostic segmentation, which is quite sensible to segment boundaries (Kirillov et al. 2023) and has been investigated into various domains. Specifically, SAM operates in a manner that requires a preexisting prompt such as a reference point, bounding box, or mask, to accompany the input image. Obviously, category-agnostic segmentation provided by SAM cannot be directly applied to semantic segmentation. Hence, several studies explore refined manual prompts for category-specific segmentation in domain-specific applications, such as manually-labeled bounding boxes for medical image segmentation (Wu et al. 2023), showcasing promising results. Therefore, considering the limitation of blurry boundary recognition in existing urban village identification studies as well as the strength of generalization and boundary sensitivity in SAM, an interesting research question is whether SAM can help urban village identification from satellite images.

Regarding the research question above, in this paper, we propose a generalist-specialist-like framework called UV-SAM, to adapt SAM for urban village identification. Specifically, the critical point of adaption lies in generating category-specific prompts that can encourage SAM to focus on urban villages in satellite images. Therefore, we regard SAM with large frozen parameters as a generalist in category-agnostic segmentation and develop a semantic segmentation model with limited learnable parameters as a specialist for urban village identification, where the specialist automatically generates prompts for the generalist while the outputs by the generalist in turn update model parameters of the specialist. Following the proposed framework, UV-SAM employs four distinct categories of prompts specifically for urban villages in satellite images. Firstly, UV-SAM develops a small-sized semantic segmentation model like SegFormer (Xie et al. 2021) to produce coarse segmentation masks for urban villages, based on which mask prompts and box prompts of urban villages are generated. Secondly, the feature maps from image encoders in both SAM and SegFormer are extracted as semantic prompts. Furthermore, a prompt mixer module is designed to fuse such four types of prompts together, and the resulting urban village prompt vector is fed into SAM for urban village specific segmentation. In summary, our contributions lie in three aspects:

- We are the first to introduce the vision foundation model SAM for urban village identification, which enlightens the application of foundation models in artificial intelligence for sustainable cities and SDG.
- We establish a novel generalist-specialist framework, UV-SAM, which automatically generates four distinct types of prompts, and seamlessly integrates SAM into urban village identification applications.
- We conduct extensive experiments on two cities Beijing and Xi'an in China, and the results demonstrate that our proposed framework achieves significant performance improvement compared with state-of-the-art models. Further case studies reveal the evolving trends of urban villages in both amount and area, as well as their spatial distribution, which provides valuable insights for urban planning and governance.

Related Works

Satellite Image-based Urban Village Identification. Urban village identification refers to the process of identifying areas or regions within a city that exhibit characteristics of urban villages, which are crucial for understanding the spatial distribution and evolution of urban villages.

Several studies investigate the satellite image classification problem to identify whether urban villages exist in corresponding images. Earlier studies (Huang, Liu, and Zhang 2015; Liu et al. 2017) apply traditional machine learning algorithms, such as support vector machines, to classify urban and non-urban areas based on handcrafted features. Recent studies employ deep learning techniques, particularly convolutional neural networks (CNN), to automatically learn discriminative features from satellite images. For example, some studies (Chen et al. 2022; Fan et al. 2022a) classify urban villages by constructing various deep learning models over satellite images and street images. Another study (Fan et al. 2022b) classifies urban informal settlements using very high-resolution remote sensing images and time-series population density data. Also a recent work (Xiao et al. 2023) uses an urban region graph and designs a contextual master-slave framework to effectively detect the urban village. However, these studies focus on image classification and fail to identify urban village boundaries, providing limited information for sustainable cities.

On the other hand, some studies formulate urban village identification as a segmentation problem. For example, the Mask R-CNN model is used to detect urban villages and segment the boundaries of urban villages from satellite images (Chen et al. 2019). Another two studies (Mast, Wei, and Wurm 2020; Pan et al. 2020) respectively utilize the well-established semantic segmentation models, including Fully Convolutional Neural Networks (FCN) and U-Net, to map urban village areas in Shenzhen and Guangzhou. Moreover, UVLens (Chen et al. 2021) employs taxi trajectories to divide the city-wide satellite image into smaller patches and then incorporates bike-sharing drop-off data into these image patches and utilizes the Mask R-CNN model (He et al. 2017) to detect urban villages. Overall, existing studies on urban villages either focus narrowly on classification or struggle with inaccurate semantic segmentation. Besides, such studies often rely on additional data sources such as street views and traffic data, which do not apply to all cities.

SAM Applications. Since the proposal in April 2023, SAM has been widely used in different fields, such as medical image processing (Ma and Wang 2023; Zhou et al. 2023), 3D vision (Cen et al. 2023; Shen, Yang, and Wang 2023), inpainting (Yu et al. 2023), object tracking (Yang et al. 2023; Rajič et al. 2023) and so on, which fall into two application ways: (i) Fine-tuning or adding an adapter on SAM im-

age encoder. For example, SAMed (Zhang and Liu 2023), MedSAM (Wu et al. 2023) and 3DSAMadpter (Gong et al. 2023) entail the customization of SAM specifically for medical image segmentation with adapters incorporated, yielding performance improvement in medical image segmentation tasks. (ii) Generating task-specific prompts. For example, AutoSAM (Shaharabany et al. 2023) designs an auxiliary convolution network that replaces the prompt embedding for medical imaging domains. RSPrompter (Chen et al. 2023) develops anchor-based and query-based prompts with SAM for satellite image-based instance segmentation. Motivated by such SAM-based applications, we adapt SAM to the urban village identification problem.

Preliminary

In this section, we provide problem statement and important models of SegFormer and SAM used in the methodology.

Problem Statement. Urban village identification refers to the task of identifying and delineating the boundaries of urban villages within a given geographical area, separating them from the surrounding areas. Thus, the research problem with satellite images is formally defined as:

Problem 1 *Given any satellite image $\mathcal{I}$, the satellite image-based urban village identification problem is to design method f to identify specific boundaries U for urban village therein (if existed), denoted as $U: f(I) \rightarrow U$.*

SegFormer. SegFormer (Xie et al. 2021) builds an encoder-decoder framework to achieve impressive performance in semantic segmentation tasks. In the encoder part, SegFormer employs a hierarchical pyramid Vision Transformer (ViT) (Dosovitskiy et al. 2020) to break down the input image into hierarchical regions and process them at different levels of abstraction. In the decoder part, a multi-layer perceptron (MLP) is developed to gather information from various layers, effectively merging local attention and global attention mechanisms to create potent representations, which are finally upsampled to produce the ultimate segmentation mask.

SAM. SAM (Kirillov et al. 2023) designs a flexible prompting-enabled model architecture for category-agnostic segmentation. To be specific, SAM consists of an image encoder, a prompt encoder, and a mask decoder, where the image encoder is pre-trained using the masked autoencoder technique, the prompt encoder handles dense and sparse inputs, and the mask decoder predicts the masks based on the encoded embeddings. Especially, SAM supports external prompts like boxes, points, and texts for segmenting objects.

Methodology

Generalist-Specialist Framework Overview. Figure 2 presents the main framework of our proposed UV-SAM model for the urban village identification problem, which falls into a generalist-specialist style. Considering the limitations of existing models for urban village identification, which struggle to accurately define the boundaries of urban villages, in the generalist part, we leverage SAM's robust edge detection capabilities to learn such finer boundaries. Moreover, in the specialist part, to provide urban village-specific prompts for SAM, we employ a lightweight semantic segmentation model, SegFormer, for prompt generation.

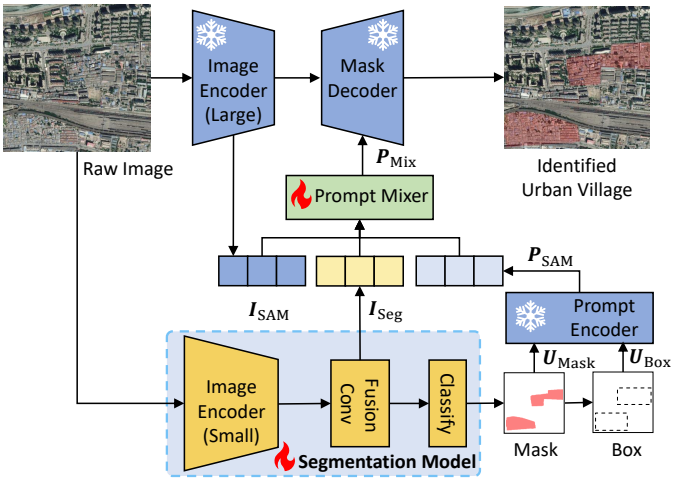


Figure 2: The illustration of proposed UV-SAM framework. The snowflake and torch symbols in the figure signify that the model parameters in this part are kept frozen and learnable, respectively.

Image Encoder. Specifically, given a satellite image I as input, the image is fed into SAM's image encoder $\Phi_{\text{SAM-Image}}$ with large pre-trained parameters and SegFormer's image encoder $\Phi_{\text{Seg-Image}}$ with small learnable parameters, with output as $\boldsymbol{I}_{\text{SAM}}$ and $\boldsymbol{I}_{\text{Seg}}$, respectively. Since $\boldsymbol{I}_{\text{Seg}}$ comprises features of multiple scales, UV-SAM applies a fusion layer of MLP Φ_{Agg} to aggregate such features. The above processes are expressed as:

$$\boldsymbol{I}_{\text{SAM}} = \Phi_{\text{SAM-Image}}(I) \tag{1}$$

$$\boldsymbol{I}_{\text{Seg}} = \Phi_{\text{Agg}}(\Phi_{\text{Seg-Image}}(I)) \tag{2}$$

Subsequently, the $\boldsymbol{I}_{\text{Seg}}$ goes through the classification layer within SegFormer to generate masks U_{Mask} for urban villages therein. After undergoing image processing, the masks are then used to derive the corresponding bounding boxes U_{Box}. This process can be expressed as:

$$U_{\text{Mask}} = \text{MLP}(\boldsymbol{I}_{\text{Seg}}) \tag{3}$$

$$U_{\text{Box}} \underset{\text{Image Process}}{\longleftarrow} U_{\text{Mask}} \tag{4}$$

Prompt Generation. The above masks and bounding boxes are utilized as prompts, and fed into the prompt encoder $\Phi_{\text{SAM-Prompt}}$ resulting in a sparse prompt embedding $\boldsymbol{P}_{\text{SAM}}$, which encompasses explicit prompt information on urban village location details, which is expressed as:

$$\boldsymbol{P}_{\text{SAM}} = \Phi_{\text{SAM-Prompt}}(U_{\text{Box}}, U_{\text{Mask}}) \tag{5}$$

Moreover, since both $\boldsymbol{I}_{\text{SAM}}$ and $\boldsymbol{I}_{\text{Seg}}$ aggregate abstract semantic information specific to urban villages, UV-SAM also models both as semantic prompts. Especially, a prompt generation module is designed to fuse such four types of

prompts together, whose two prompt mixing variants of $\Phi_{\text{Mix}}^{\text{Add}}$ and $\Phi_{\text{Mix}}^{\text{MLP}}$ can be expressed as:

$$\boldsymbol{P}_{\text{Mix}}=\Phi_{\text{Mix}}^{\text{Add}}(\boldsymbol{P}_{\text{SAM}}, \boldsymbol{I}_{\text{Seg}}, \boldsymbol{I}_{\text{SAM}})=\boldsymbol{P}_{\text{SAM}} + \boldsymbol{I}_{\text{Seg}} + \boldsymbol{I}_{\text{SAM}} \quad (6)$$

$$\boldsymbol{P}_{\text{Mix}}=\Phi_{\text{Mix}}^{\text{MLP}}(\boldsymbol{P}_{\text{SAM}}, \boldsymbol{I}_{\text{Seg}}, \boldsymbol{I}_{\text{SAM}})=\text{MLP}([\boldsymbol{P}_{\text{SAM}}; \boldsymbol{I}_{\text{Seg}}; \boldsymbol{I}_{\text{SAM}}]) \quad (7)$$

where $\Phi_{\text{Mix}}^{\text{Add}}$ involves directly adding normalized features and ensures a straightforward fusion of insights, while $\Phi_{\text{Mix}}^{\text{MLP}}$ entails concatenating normalized features and then passing them through a projection head for dimensionality reduction, offering a more intricate yet controlled method of knowledge synthesis. The choice between the two forms could depend on the nature of the information being merged and the specific requirements of the task.

Mask Decoder. Finally, based on the mixed prompt and pre-trained mask decoder in SAM, UV-SAM identifies urban villages in satellite images as follows:

$$U = \Phi_{\text{SAM-Mask}}(\boldsymbol{I}_{\text{SAM}}, \boldsymbol{P}_{\text{Mix}}) \quad (8)$$

where elements in U identify whether specific pixels belong to urban villages.

Training Loss. Similar to SAM, in the context of the larger model, we adopt a mask prediction strategy involving a linear combination of focal loss $\mathcal{L}_{\text{focal}}$ (Lin et al. 2017), dice loss $\mathcal{L}_{\text{dice}}$ (Milletari, Navab, and Ahmadi 2016) and mean-square-error loss $\mathcal{L}_{\text{mse}}$ at a weight of 1:1:1. In addition, SegFormer continues to utilize the straightforward cross-entropy loss $\mathcal{L}_{\text{Seg}}$ for its loss function. Consequently, the overall loss can be expressed as follows:

$$\mathcal{L}_{\text{SAM}} = \mathcal{L}_{\text{focal}} + \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{mse}} \quad (9)$$

$$\mathcal{L} = \lambda\mathcal{L}_{\text{SAM}} + \mathcal{L}_{\text{Seg}} \quad (10)$$

where λ is a hyper-parameter to weigh the impacts of generalist and specialist modules.

Experiments

In this section, we conduct experiments to answer the following research questions:

- **RQ1:** How does our proposed UV-SAM model perform compared with existing baseline approaches?
- **RQ2:** What is the effectiveness of each designed module in our proposed UV-SAM model?
- **RQ3:** Can our proposed UV-SAM model identify the spatial distribution of urban villages?
- **RQ4:** Can our proposed UV-SAM model identify the evolving trends of urban villages with area and amount?

Experiment Setups

Datasets. For our research, we collected datasets consisting of satellite images from two major cities of Beijing and Xi'an in China. Table 1 reports basic statistics for datasets. For training purposes, we randomly split the dataset into three subsets of training, validation, and testing sets with the proportion of 6:2:2.

City	#Satellite Image	#Urban Village	Year
Beijing	2,491	545	2016
Xi'an	837	205	2018

Table 1: Dataset statistics.

We specifically focus on main urban areas to capture the dynamics of urban village evolution. The satellite images are obtained from ArcGIS[1] and have a resolution of approximately 1.05 meters per pixel. To prepare the dataset for training and evaluation, we merge the individual satellite images into larger images of 1024×1024 pixels. This merging process ensures that the resulting images contained comprehensive main urban area information.

As for labels, to begin with, we recruited a group of participants from urban research and provided them with appropriate incentives. We also conducted training sessions to ensure that the participants had a good understanding of urban villages and related background knowledge. To facilitate the labeling process, we utilized the EasyData[2] crowdsourcing platform, through which we randomly assigned image patches to the participants for cross-validation. Each patch was assigned to three participants to ensure accuracy and consistency in the labeling process. To maintain quality control, we assigned specific individuals to validate the mask annotations provided by the participants. This validation process helped to ensure the accuracy and reliability of the obtained mask labels. By conducting per-pixel voting, we obtained the ground truth labels for all the image patches.

Baselines. We have conducted a comparative analysis of our model against various baseline approaches:

- **FCN** (Long, Shelhamer, and Darrell 2015). FCN replaces fully connected layers with convolutional layers, enabling end-to-end pixel-wise predictions.
- **DeepLabv3+** (Chen et al. 2018). The architecture of DeepLabv3+ is enhanced by integrating atrous spatial pyramid pooling and decoder modules, resulting in a more sophisticated design.
- **UVLens** (Chen et al. 2021). UVLens integrates bike-sharing drop-off data and satellite images into image patches and applies Mask R-CNN (He et al. 2017) model for urban village identification.
- **RSPrompter** (Chen et al. 2023). RSPrompter incorporates elements from both Faster R-CNN (He et al. 2016a) and Transformer (Vaswani et al. 2017) architectures into the prompt generation process for satellite image instance segmentation.

Metrics. To evaluate the accuracy of our identification method, we compare the segmented urban villages with the ground truth dataset under two types of metrics in respect to detection accuracy and segmentation accuracy.

For detection accuracy, if a detected urban village spatially overlaps with an urban village in the ground-truth

[1]https://geoenrich.arcgis.com/

[2]https://ai.baidu.com/easydata/

Dataset	Beijing				Xi'an			
Method	IoU	F1-Score	Recall	Precision	IoU	F1-Score	Recall	Precision
FCN	0.660	0.802	0.752	0.859	0.720	0.833	0.800	0.870
DeepLabv3+	0.650	0.787	0.719	**0.870**	0.668	0.821	0.780	0.867
UVLens	0.623	0.783	0.777	0.790	0.687	0.863	0.880	0.867
RSPrompter	0.462	0.687	0.860	0.571	0.568	0.800	0.800	0.800
UV-SAM	**0.721**	**0.871**	**0.893**	0.851	**0.747**	**0.904**	**0.940**	**0.871**

Table 2: Overall performance of UV-SAM and baselines on two datasets. Bold denotes the best results and underline denotes the second-best results.

dataset, we mark it as true positive, otherwise false positive, based on which we calculate precision, recall and F1-score.

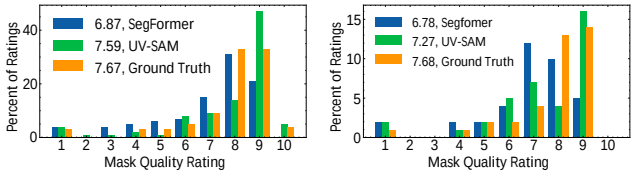
For segmentation accuracy, we utilize the widely-used Intersection over Union (IoU) metric, which is calculated as the intersection area divided by the union area between the segmented urban villages and the corresponding ground-truth urban villages.

Implementation. In our experiments, we consistently employ the ViT-Large backbone of SAM and the MiT-B0 lightweight encoder of SegFormer, unless specified otherwise. We select the Adam optimizer to facilitate parameter learning and incorporate a cosine annealing scheduler to gradually decrease the learning rate. The mini-batch size is fixed at 4, and the complete training process spans 100 epochs. We conduct a grid search for optimal values of the learning rate, weight decay, and λ, from $\{0.005, 0.0005, 0.00005\}$, $\{0.01, 0.001\}$ and $\{0.1, 1, 10\}$, respectively. Besides, based on the validation performance, we select $\Phi_{\text{Mix}}^{\text{MLP}}$ and $\Phi_{\text{Mix}}^{\text{Add}}$ for Beijing and Xi'an, respectively. The experiment details are available at the link[3].

Overall Performance (RQ1)

Table 2 shows the overall performance comparison on Beijing and Xi'an datasets. From these results, we have the following observations:

- **UV-SAM achieves the best performance across both datasets.** The results showcase that our proposed model achieves state-of-the-art performance, which successfully adapts SAM into urban village identification. For the segmentation accuracy, compared with the baselines, our model outperforms the best baseline by 4%-9% on IoU in two datasets. Similar performance improvements can be also found in F1-score for the detection accuracy. It is notable that performance difference on two datasets with DeepLabv3+. This is due to how well DeepLabv3+'s structure matches the characteristics of the Beijing dataset. The DeepLabv3+ architecture combines high-level features for semantic information with low-level features for capturing boundary details. Beijing's unique features, with denser traditional courtyard style housing and shorter buildings, differ from the high-rise, dense buildings of Xi'an. In addition, owing to the generalized generalist-specialist framework, all baselines in Table 2 can be incorporated into UV-SAM as the specialist module, which can bring further performance improvement for urban village identification.
- **Existing SAM-based models perform poorly in urban village identification.** According to the results in Table 2, the performance of RSPrompter (Chen et al. 2023) notably lags behind that of other baseline models in terms of IoU and F1-score, e.g., the worst IoU and F1-score of 0.462 and 0.687 in Bejing dataset. Such results suggest that the learnable prompts in RSPrompter fail to capture the intricate and abstract semantic features that hold particular relevance to urban villages, and thus provide useless guidance to SAM. Besides, the performance drop also emphasizes the non-tricky adaption of SAM to urban village identification.
- **Transformer-based encoders demonstrate a better semantic understanding of urban villages.** In the capacity of transformer-based architectures, our proposed UV-SAM exhibits remarkable superiority compared with other CNN-based models, in terms of IoU and F1-score metrics. As described before, urban villages embody intricate and advanced semantic concepts, and the demarcation of their boundaries within satellite imagery is notably influenced by contextual factors in their surrounding environment. Thus, the Transformer architecture with the attention mechanism can better capture fine-grained features therein, while CNN-based models mainly grasp higher-level semantic abstractions, leading to inaccurate boundaries and inferior performance.



(a) Beijing (b) Xi'an

Figure 3: Mask quality rating distributions by datasets from our human evaluation study in Beijing and Xi'an, with average scores shown in the legend.

Furthermore, to better evaluate the mask quality in ur-

[3]https://github.com/tsinghua-fib-lab/UV-SAM

Dataset	Beijing				Xi'an			
Variants	IoU	F1-Score	Recall	Precision	IoU	F1-Score	Recall	Precision
w/o Box	0.708	0.838	0.876	0.803	0.173	0.267	0.160	0.800
w/o Mask	0.635	0.842	0.860	0.825	0.744	0.876	0.920	0.836
w/o SAM emb	0.697	0.846	0.909	0.791	0.717	0.826	0.900	0.763
w/o Seg emb	0.694	0.832	0.901	0.773	0.733	0.849	0.900	0.804
w/o SAM	0.688	0.854	0.893	0.818	0.731	0.860	0.860	0.860
UV-SAM	**0.721**	**0.871**	**0.893**	**0.851**	**0.747**	**0.904**	**0.940**	**0.871**

Table 3: Ablation study of UV-SAM variants on two datasets.

ban villages, we introduce the human study for evaluation (Kirillov et al. 2023). Specifically, we present masks generated by models to annotators and require them to rate the quality of each mask from 1 to 10. A score of 1 means that the mask has no relevance to urban villages while 10 indicates that there are no noticeable errors in the identified boundaries of urban village areas. We conduct a comparison between the masks produced by SAM and those generated by SegFormer, along with the ground truth data, which are presented in Figure 3. The results show that **UV-SAM achieves better mask quality than SegFormer across both datasets**. For example, on the Beijing dataset, UV-SAM achieves an average rating of 7.59, while SegFormer is 6.87, compared with the ground truth of 7.67. Within the lower score range, UV-SAM's performance falls short of the baseline counterpart. Conversely, in the higher score range, there is a conspicuous increase in frequency. Such results demonstrate the effectiveness of SAM for segmenting boundaries.

Ablation Study (RQ2)

To evaluate the effectiveness of each module in UV-SAM, Table 3 shows the detection and segmentation performance of different model variants on both datasets. According to the results, without the box prompt, our model performance drops 1.8% and 75.9%. Thus, the box prompt plays an important role in the performance guarantee, which guides the mask decoder of the SAM to focus on the regions of interest. Besides, the performance drop on Xi'an dataset can be largely attributed to the straightforward way of prompt addition. Moreover, the mask prompt offers a dense embedding that specifically emphasizes the boundaries of objects within the image, contributing 11.9% and 0.4% on IoU for two datasets, respectively. Furthermore, with abstract semantic information specific to urban villages, the SAM embedding from the image encoder (large) further achieves 3%-4% improvement on IoU. Finally, our model can get an improvement of 2%-3% with the assistance of segmentation embedding obtained from the image encoder(small), capturing the high-level semantics provided by the specialist-like semantic segmentation model. Thus, all four types of prompts are essential for effective urban village identification. In addition, our model performance drops 4.6% and 2.1% without the SAM. So a generalist-like SAM provides more accurate boundary information for urban village identification.

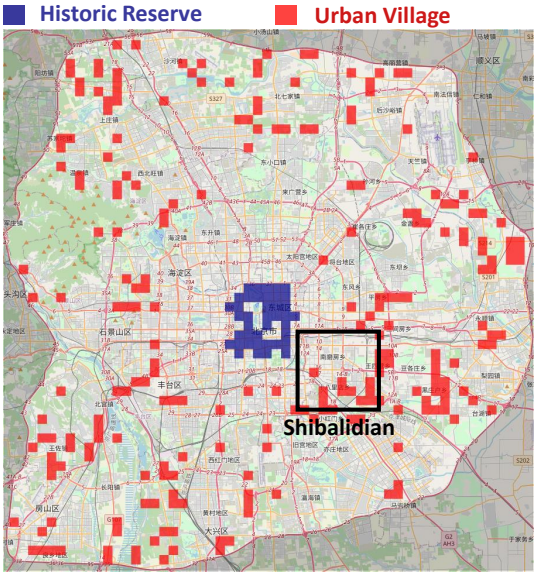


Figure 4: Urban village (UV) distribution in Beijing in 2020.

Spatial Distribution Analysis (RQ3)

To mitigate the potential risks of urban villages to urban development as well as improve citizens' living conditions, the governments often initiate gradual demolitions and resident relocations therein. Thus, identifying the spatial distribution of urban villages offers a crucial reference point for urban planning. In Figure 4, we visualize the spatial distribution of urban villages within the sixth ring road of Beijing in 2020.

As depicted in the figure, within the second ring road of Beijing, numerous historical reserve blocks are presented, which often consist of courtyard-style housing, accommodating a few households in close proximity. Despite their historical and conservation value, these areas typically have small per capita living spaces, poor sanitation conditions, and low greenery coverage, which align well with the definition of urban villages. On the contrary, urban villages are found to be distributed more thinly between the third and sixth ring roads. Particularly, there is a noticeable clustering of urban villages near Shibalidian Township[4], which is a famous urban village cluster in Beijing. Moreover, the south-

[4]https://en.wikipedia.org/wiki/Shibalidian

ern and eastern parts of this area show a higher density of urban villages compared to the western and northern parts. This discrepancy in distribution could be attributed to historical population movement patterns and local levels of economic development.

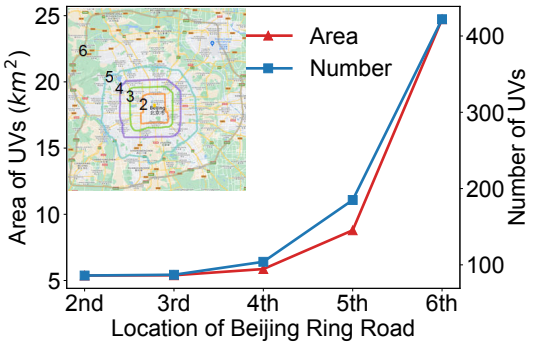
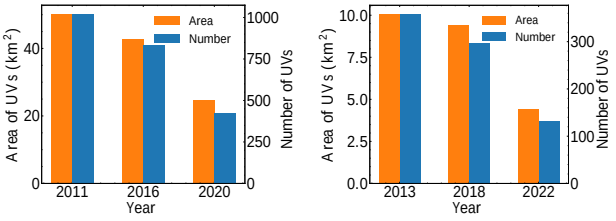


Figure 5: Urban village (UV) distribution along Beijing's ring roads with respect to area and amount in 2020.

To quantify the spatial distribution, we further plot the urban village distribution curves along Beijing's ring roads with respect to area and amount in Figure 5. We crudely determined the count and extent of urban villages by utilizing the number of predicted masks and their cumulative pixel values derived from satellite imagery results. According to the results, there is a significant increase in both area and number of urban villages between the fifth and sixth ring roads, where the distance to the urban center is far enough and the buildings from the original village are preserved.



(a) Beijing (b) Xi'an

Figure 6: The change of the area and number of urban villages over years in Beijing and Xi'an.

Evolving Trend (RQ4)

To understand the formation, expansion and shrinkage of urban villages, in Figure 7, we illustrate the variations in urban village area and quantity between different years for the cities of Beijing and Xi'an, using the satellite images captured at various time points.

According to the results, the city of Beijing hosted an estimated 1,000 urban villages in 2011, and Xi'an accommodated around 360 urban villages in the year 2013. By 2016 or 2018, the urban villages only decreased by less than 10 square kilometers. However, a notable transformation occurred by the year 2020. During this time, both the spatial extent and numerical prevalence of urban villages underwent a remarkably rapid contraction, resulting in a reduction of fifty percent compared to their previous levels. This discernible trend is plausibly attributable to the promulgation of the Beijing Urban Master Plan(2016-2035) by the governmental authorities.

Especially, the example of Jijiamiao Village serves as a case currently undergoing transformation[5]. As shown in the Figure 7, the Jijiamiao Village is surrounded by high-rise buildings, their outdated structures no longer in sync with the modern landscape. As early in 2011, policies were introduced to gradually renovate these aging structures. Therefore, by 2016, their presence had diminished compared to 2011. By 2020, they had nearly disappeared entirely. The surrounding green spaces and high-rise buildings are also undergoing slow but steady development.

(a) The Jijiamiao Village in 2011 (b) The Jijiamiao Village in 2016 (c) The Jijiamiao Village in 2020

Figure 7: The boundary change of the Jijiamiao Village from 2011 to 2020. The highlighted red regions denote the areas where urban villages have been identified, signifying a gradual reduction over time.

Conclusion

In this paper, we propose UV-SAM, a vision foundation model-based framework for urban village identification. The UV-SAM framework introduces a specialist-like semantic segmentation model to generate four types of urban village-specific prompts and then feeds into a generalist-like SAM model to identify urban village boundaries from satellite images. Through comprehensive experiments, we substantiate the effectiveness of our model across various datasets, which also provide deep insights into the spatial distributions and temporal trends of urban villages. Moreover, our study demonstrates the possibility of vision foundation models for sustainable development goals and sustainable cities.

Despite surpassing baseline performance, it's noteworthy that our results may exhibit a certain degree of reduced interpretability. Thus, in future work, we aim to delve into the intricate interplay of features that underlie the emergence and dissolution of urban villages. We also plan to transfer the proposed framework to slum identification in cities and help understand the global informal settlements.

[5]http://zjw.beijing.gov.cn/bjjs/gcjs/zdgcjs/2016/xmjh/363391/index.shtml

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under 2022YFF0606904, the National Natural Science Foundation of China under U22B2057 and U21B2036.

EXPERIMENT DETAILS

Experiment Details for Spatial Distributions and Evolving Trends

Here we introduce the details of the pre-classification module for spatial distributions and evolving trends. Due to the limited size of the dataset, we introduced a binary classification model specifically to distinguish urban and non-urban areas before the semantic segmentation process. The module is only applied in the spatial analysis of specific cities and is independent of our UV-SAM framework.

Datasets. For the training classification model, we construct a dataset for each specific city. The urban village area is used as positive samples and the non-urban village areas are randomly selected as negative samples, ensuring a near 1:1 positive-negative ratio in the training, validation, and test sets.

Implementation. ResNet50 (He et al. 2016b) is used to implement the classification model. We select the Adam optimizer to facilitate parameter learning and incorporate a cosine annealing scheduler to gradually decrease the learning rate. The learning rate is set as 0.0001, and the batch size is fixed at 32. To quantitatively measure the performance of the classification model, we adopt the Area Under Curve (AUC), Recall, Precision, and F1-score as evaluation metrics.

Dataset	AUC	F1-Score	Precision	Recall
Beijing	0.964	0.868	1.00	0.767
Xi'an	0.756	0.677	0.786	0.595

Table 4: Performance of classification module on two datasets.

Performance. Table 4 shows the performance of Beijing and Xi'an datasets. We note a considerable level of precision achieved in the Beijing dataset, whereas the Xi'an dataset demonstrates noticeably lower precision. This substantial contrast in results can be attributed to a combination of key factors, notably including the limited scope of the Xi'an dataset and the inherent differences in the quality of available satellite imagery.

Spatial Distribution Analysis with Street View Images in Beijing

Due to the high cost of collecting street view imagery, obtaining street view images for any given time and location is not feasible. Consequently, we can only employ such imagery as an auxiliary tool to facilitate the observation of urban village evolution.

As illustrated in Figure 8, we choose three street view images from both the historical reserve area and the urban village cluster of Shibalidian, respectively. These images are intended to showcase three distinctly different styles of urban village environments.

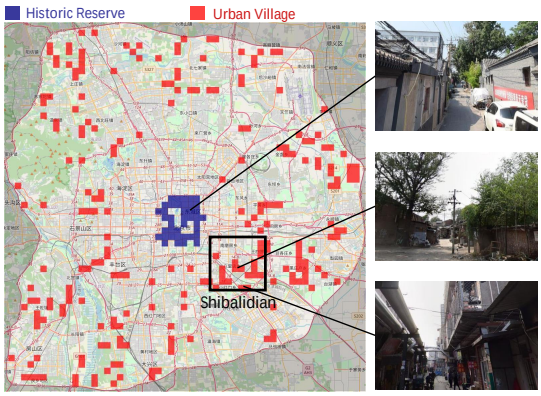
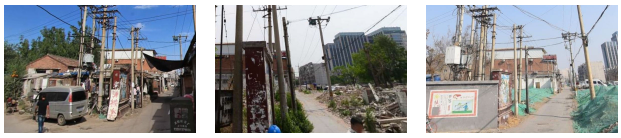


Figure 8: Urban village distribution with street view images in Beijing in 2020.



(a) The Jijiamiao Village in 2013 (b) The Jijiamiao Village in 2015 (c) The Jijiamiao Village in 2019

Figure 9: The environment change of the Jijiamiao Village from 2013 to 2019 at street view imagery.

Evolving Trend with Street View Images in Beijing

As shown in Figure 9, We show three street view images at the same location in different years. Compared to the chaotic condition of the urban village in 2013, the village had been demolished in 2015. In 2019, new walls were built to improve the city's appearance.

Spatial Distribution Analysis with Street View Images in Xi'an

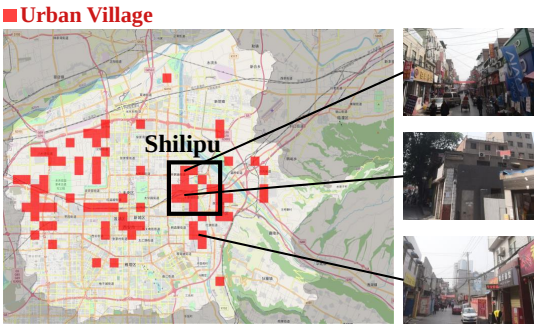


Figure 10: Urban village distribution in Xi'an in 2022.

In Figure 10, we visualize the spatial distribution of urban villages within the urban areas of Xi'an, including Lianhu District, Xincheng District, Beilin District, Yanta District, Baqiao District and Weiyang District. In Xi'an, there are few historical reserve blocks. As shown in the figure, the urban villages within the main urban areas demonstrate a distinct pattern of horizontal distribution. Notably, in the eastern area, a distinct clustering of urban villages is evident in close proximity to Shilipu Village. As for the central area of Xi'an, several urban villages are identified, which might be related to potential misidentification caused by the lower quality of satellite imagery.

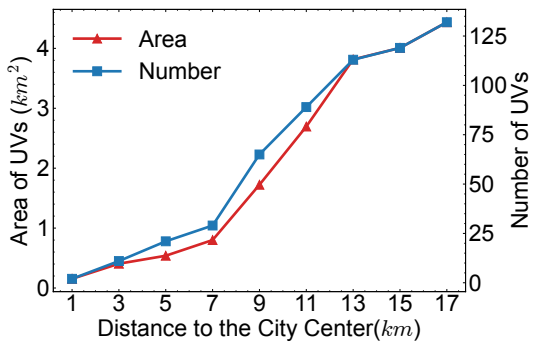


Figure 11: Urban village (UV) distribution in Xi'an with respect to area and amount in 2022.

Furthermore, we plot distribution curves to visually depict the spatial distribution patterns. In Figure 11, the area and amount of urban villages in 2022 are showcased in relation to their distances from the city center of Xi'an. When the distance from the city center is between 7 to 11 kilometers, there is a steep linear increase in the number and area of urban villages. But at larger distances, the rates of increase become less pronounced. This can be attributed to the "urban-suburban-rural" structure created by the rapid process of urban expansion, where a large number of urban villages are concentrated in the suburbs.

Evolving Trend in Xi'an

(a) The Yangjia Village in 2013 (b) The Yangjia Village in 2018 (c) The Yangjia Village in 2022

Figure 12: The boundary change of the Yangjia Village from 2013 to 2022. The highlighted red regions denote the areas where urban villages have been identified, signifying a gradual reduction over time.

The Yangjia Village, where the historical ruin of the Mingde Gate is located, became a popular choice for temporary residents due to its affordable rental options. In 2013, the government had planned to take down and renovate the

village[6]. As shown in Figure 12, the majority of the Yangjia Village was deconstructed in 2018. In 2022, a small part of the Yangjia Village had been transformed into high-rise buildings, while the majority part had been developed into a heritage park.

(a) The Yangjia Village in 2014

(b) The Yangjia Village in 2017

(c) The Yangjia Village in 2019

Figure 13: The environment change of the Yangjia Village from 2014 to 2019.

From the street view images of Yangjia Village displayed in Figure 13, it becomes clear that the village started being renovated as early as 2014. By 2019, the emergence of tall buildings became noticeable.

References

Cen, J.; Zhou, Z.; Fang, J.; Shen, W.; Xie, L.; Zhang, X.; and Tian, Q. 2023. Segment Anything in 3d with Nerfs. *arXiv preprint arXiv:2304.12308*.

Chen, B.; Feng, Q.; Niu, B.; Yan, F.; Gao, B.; Yang, J.; Gong, J.; and Liu, J. 2022. Multi-modal Fusion of Satellite and Street-view Images for Urban Village Classification Based on a Dual-branch Deep Neural Network. *International Journal of Applied Earth Observation and Geoinformation*, 109: 102794.

Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; and Shi, Z. 2023. RSPrompter: Learning to Prompt for Remote Sensing Instance Segmentation Based on Visual Foundation Model. *arXiv preprint arXiv:2306.16269*.

Chen, L.; Lu, C.; Yuan, F.; Jiang, Z.; Wang, L.; Zhang, D.; Luo, R.; Fan, X.; and Wang, C. 2021. UVLens: Urban Village Boundary Identification and Population Estimation Leveraging Open Government Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2): 1–26.

Chen, L.; Xie, T.; Wang, X.; and Wang, C. 2019. Identifying Urban Villages from City-wide Satellite Imagery Leveraging Mask R-CNN. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 29–32.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European Conference on Computer Vision*, 801–818.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*.

Fan, R.; Li, J.; Li, F.; Han, W.; and Wang, L. 2022a. Multilevel Spatial-Channel Feature Fusion Network for Urban Village Classification by Fusing Satellite and Streetview Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13.

Fan, R.; Li, J.; Song, W.; Han, W.; Yan, J.; and Wang, L. 2022b. Urban Informal Settlements Classification via a Transformer-based Spatial-temporal Fusion Network Using Multimodal Remote Sensing and Time-series Human Activity Data. *International Journal of Applied Earth Observation and Geoinformation*, 111: 102831.

Gong, S.; Zhong, Y.; Ma, W.; Li, J.; Wang, Z.; Zhang, J.; Heng, P.-A.; and Dou, Q. 2023. 3DSAM-adapter: Holistic Adaptation of SAM from 2D to 3D for Promptable Medical Image Segmentation. *arXiv preprint arXiv:2306.13465*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Identity Mappings in Deep Residual Networks. In *Proceedings of the European Conference on Computer Vision*, 630–645.

Huang, X.; Liu, H.; and Zhang, L. 2015. Spatiotemporal Detection and Analysis of Urban Villages in Mega City Regions of China Using High-resolution Remotely Sensed Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7): 3639–3657.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment Anything. *arXiv preprint arXiv:2304.02643*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Liu, H.; Huang, X.; Wen, D.; and Li, J. 2017. The Use of Landscape Metrics and Transfer Learning to Explore Urban Villages in China. *Remote Sensing*, 9(4): 365.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Ma, J.; and Wang, B. 2023. Segment Anything in Medical Images. *arXiv preprint arXiv:2304.12306*.

Mast, J.; Wei, C.; and Wurm, M. 2020. Mapping Urban Villages Using Fully Convolutional Neural Networks. *Remote Sensing Letters*, 11(7): 630–639.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *2016 Fourth International Conference on 3D Vision*, 565–571. IEEE.

[6]https://www.xa.gov.cn/gk/zcfg/xaszfwj/5d49119d65cbd87465a7851d.html

Nath, D. 2016. Making Cities and Human Settlements Inclusive, Safe, Resilient and Sustainable, sl: Sustainable Urbanization Summit.

Pan, Z.; Xu, J.; Guo, Y.; Hu, Y.; and Wang, G. 2020. Deep Learning Segmentation and Classification for Urban Village Using a Worldview Satellite Image Based on U-Net. *Remote Sensing*, 12(10): 1574.

Rajič, F.; Ke, L.; Tai, Y.-W.; Tang, C.-K.; Danelljan, M.; and Yu, F. 2023. Segment Anything Meets Point Tracking. *arXiv preprint arXiv:2307.01197*.

Shaharabany, T.; Dahan, A.; Giryes, R.; and Wolf, L. 2023. AutoSAM: Adapting SAM to Medical Images by Overloading the Prompt Encoder. *arXiv preprint arXiv:2306.06370*.

Shen, Q.; Yang, X.; and Wang, X. 2023. Anything-3d: Towards Single-view Anything Reconstruction in the Wild. *arXiv preprint arXiv:2304.10261*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All You Need. *Advances in Neural Information Processing Systems*, 30.

Wikipedia contributors. 2023. Urban Village (China) — Wikipedia, The Free Encyclopedia. [Online; accessed 20-December-2023].

Wu, J.; Fu, R.; Fang, H.; Liu, Y.; Wang, Z.; Xu, Y.; Jin, Y.; and Arbel, T. 2023. Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. *arXiv preprint arXiv:2304.12620*.

Xiao, C.; Zhou, J.; Huang, J.; Zhu, H.; Xu, T.; Dou, D.; and Xiong, H. 2023. A Contextual Master-slave Framework on Urban Region Graph for Urban Village Detection. In *2023 IEEE 39th International Conference on Data Engineering*, 736–748. IEEE.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Yang, J.; Gao, M.; Li, Z.; Gao, S.; Wang, F.; and Zheng, F. 2023. Track Anything: Segment Anything Meets Videos. *arXiv preprint arXiv:2304.11968*.

Yu, T.; Feng, R.; Feng, R.; Liu, J.; Jin, X.; Zeng, W.; and Chen, Z. 2023. Inpaint Anything: Segment Anything Meets Image Inpainting. *arXiv preprint arXiv:2304.06790*.

Zhang, K.; and Liu, D. 2023. Customized Segment Anything Model for Medical Image Segmentation. *arXiv preprint arXiv:2304.13785*.

Zheng, S.; Long, F.; Fan, C. C.; and Gu, Y. 2009. Urban Villages in China: A 2008 Survey of Migrant Settlements in Beijing. *Eurasian Geography and Economics*, 50(4): 425–446.

Zhou, T.; Zhang, Y.; Zhou, Y.; Wu, Y.; and Gong, C. 2023. Can SAM Segment Polyps? *arXiv preprint arXiv:2304.07583*.