Zero-shot Building Attribute Extraction from Large-Scale Vision and Language Models

Fei Pan[1] Sangryul Jeon[2] Brian Wang[1] Frank Mckenna[3] Stella X. Yu[1,3]
[1]University of Michigan, Ann Arbor [2]Pusan National University [3]University of California, Berkeley
{feipan, bswang, stellayu}@umich.edu, srjeonn@pusan.ac.kr, fmckenna@berkeley.edu

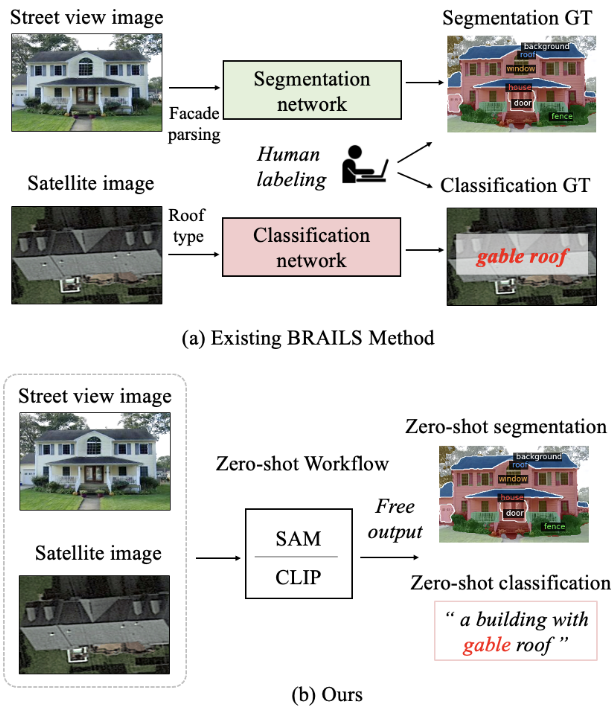
Abstract

Existing building recognition methods, exemplified by BRAILS, utilize supervised learning to extract information from satellite and street-view images for classification and segmentation. However, each task module requires human-annotated data, hindering the scalability and robustness to regional variations and annotation imbalances. In response, we propose a new zero-shot workflow for building attribute extraction that utilizes large-scale vision and language models to mitigate reliance on external annotations. The proposed workflow contains two key components: image-level captioning and segment-level captioning for the building images based on the vocabularies pertinent to structural and civil engineering. These two components generate descriptive captions by computing feature representations of the image and the vocabularies, and facilitating a semantic match between the visual and textual representations. Consequently, our framework offers a promising avenue to enhance AI-driven captioning for building attribute extraction in the structural and civil engineering domains, ultimately reducing reliance on human annotations while bolstering performance and adaptability.

1. Introduction

As a consequence of global warming, many natural hazards, such as earthquakes, hurricanes, floods, and typhoons, increase in intensity and have a destructive impact on many residential areas. To understand the impact and prepare for the potential damage caused by such hazards, the researchers in the field of natural hazards engineering and local and state agencies gather information about the buildings and other infrastructures in the areas to be studied [6]. Serving as a pivotal undertaking for building inventory generation and management, the task of building attribute extraction strives to provide information containing the number of floors, roof types, year of construction, etc.

Previous works [15,38] have utilized deep learning models to obtain building information from satellite and street view images acquired from mapping agencies. [44,45] propose to learn a model capable of identifying seismically vulnerable buildings from the street view images. These works have been collected into a software package, known as BRAILS (Building Recognition using AI at Large Scale)



(a) Existing BRAILS Method

(b) Ours

Figure 1. The existing method *vs.* ours for building attribute extraction. (a) The existing prominent method BRAILS obtains building attributes from the satellite and street view images via classification and segmentation modules which require human annotation data. (b) We propose a new zero-shot workflow to extract building attributes based on large-scale models. Our workflow uses a single module to directly extract building attributes for different tasks without human annotations and shows more robustness to novel domains.

[39], which extracts building attributes such as roof shape, building height, and foundation type from satellite and street view images by employing supervised learning in the vision domain. BRAILS provides an interface allowing building inventories to be automatically generated for a region. Moreover, it contains multiple modules and each module is formulated as either image classification, object detection, or semantic segmentation. An example of BRAILS is shown in Fig. 1a.

Nonetheless, the modules in BRAILS require data manually annotated by humans, particularly for the segmentation task which requires pixel-level annotations. These annotations could be obtained from different data providers or agencies, and detailed image-level descriptions are most often collected manually through a crowd-sourcing website. Nevertheless, there are several challenges that hinder the scalability and robustness of BRAILS.

1. The amount of human-generated annotations is inadequate to account for the large regional variations in visual and geometrical appearances, and these annotations are also prone to subjective errors.
2. Models trained on the available data struggle to effectively generalize to novel buildings in unseen regions.
3. The distribution of annotations across known classes may be biased or skewed, resulting in severe imbalance between minority classes of handful instances and majority classes of hundreds of instances, presenting a hard imbalanced classification problem.

All these factors lead to poor model generalization performance across regions.

We propose a zero-shot workflow (Fig. 1b) that tackles these challenges by utilizing large-scale models, CLIP [31] and SAM [23], which are trained on extensive and diverse datasets and can be readily adapted to downstream tasks.

- CLIP is a large-scale model that associates images with text through contrastive learning. It has the capacity to generate captions for novel images.
- SAM is a large-scale image segmentation model trained on many high-resolution images along with their annotated segmentation masks. It can provide high-quality segmentation boundaries in novel images.

Our zero-shot workflow integrates the strengths of CLIP and SAM, extracting attributes of interest to structure and civil engineers without relying on human annotations. These building attributes extracted from our workflow can be used to complete the building inventory as shown in Fig. 2.

Our workflow has two components: image-level captioning designed for image classification, and segment-level captioning for image segmentation.

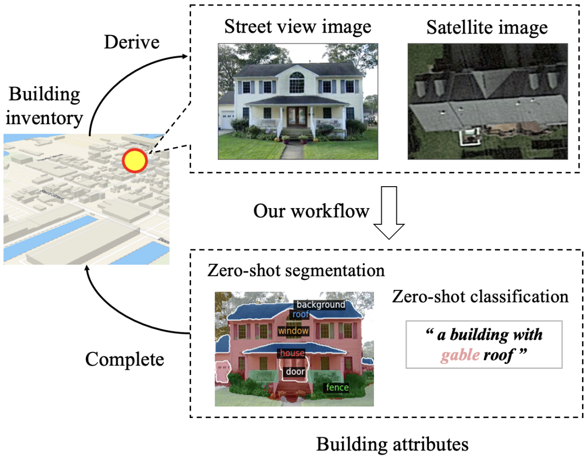


Figure 2. Our workflow directly infers building attributes on street view and satellite images. These building attributes can be used to complete the building inventory.

- In the image-level captioning component, captions are generated for building images using a list of text terms of interest to structural and civil engineers. CLIP computes feature vectors from both the image and the text terms and selects the term most similar to the image.
- The segment-level captioning component operates in a similar manner but begins by sending the building image to SAM first in order to obtain image segments. Then all these image segments are fed into CLIP to obtain a proper segment-level captioning.

Our work makes three major contributions. **1)** We are the first to utilize large-scale vision and language models for building attribute extraction in the structural and civil engineering domains. **2)** Our workflow facilitates zero-shot image classification and segmentation, applicable to any vocabulary of interest for building description, without reliance on human annotations. **3)** Our workflow harnesses the generalization capacity of CLIP for image captioning on building images, thereby enabling a robust and versatile building attribute extraction.

2. Related Work

Learning-based building recognition. Building information modeling [37] in structural and civil engineering has been employed to manage buildings and infrastructures, as well as to minimize the impact of natural hazards like hurricanes, floods, and tornadoes [6]. Prior studies [15, 38] have employed deep learning models to grasp building information from satellite and street view images sourced from map agencies. [45] and [44] introduce a model that learns to identify seismically vulnerable buildings from street view images. [11] leverages visual cues like cars in

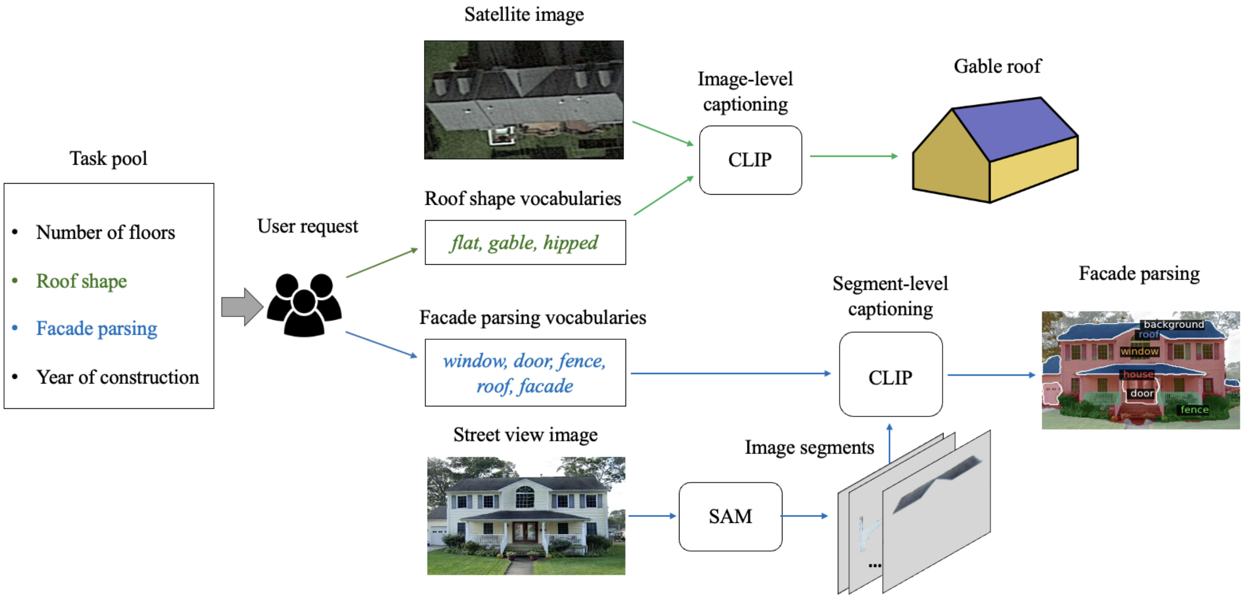


Figure 3. Our zero-shot workflow, leveraging large-scale vision and language models CLIP and SAM, effectively extracts building attributes for different tasks using the same module without human-annotated data, In contrast, traditional methods require annotated data to train multiple modules for different tasks. Our method contains two components: image-level captioning for image segmentation, and segment-level captioning for image segmentation for different tasks. Given a task requested by users, we first build a curated list of task-related vocabularies about building attributes. For the task processed by image-level captioning, CLIP generates a vocabulary from an image as input. For the task processed by segment-level captioning, CLIP predicts a vocabulary for each image segment produced by SAM. Benefiting from large-scale models, the proposed workflow shows strong generalization to the novel domain.

street view images to estimate neighborhood demographics. BRAILS [39] extracts building information from satellite and street view images using supervised learning. Each module in BRAILS is formulated as either an image classification, object detection, or semantic segmentation task. Nonetheless, the modules within BRAILS require human-annotated data, hampering the scalability and generalization of the framework.

Our approach tackles these challenges by making use of large-scale models already trained extensively on diverse data. It extracts building attributes without the need for external annotations.

Transferable features from pre-training. Feature presentations derived from models that were pre-trained on the ImageNet [8] dataset has been widely used [3, 17, 27, 32]. These representations have demonstrated the ability to generalize well [43] across various tasks such as object detection [17, 26, 32] and semantic segmentation [3, 27, 47]. Alternatively, self-supervised learning methods which involve some pretext tasks [9, 10, 28], contrastive learning [4, 5, 16], clustering [1], or bootstrapping [13], have been employed to generate versatile feature representations. Some approaches involve learning visual representation through natural language [12, 30, 34, 35], where pairs of images and corresponding captions are utilized. Recent advances such as CLIP [31] and ALIGN [21] have employed contrastive learning on extensive curated sets of image-text pairs, showcasing the pre-trained feature representations with remarkable zero-shot transfer capability. Specifically, the pre-trained CLIP model has been successfully used in style manipulation [29], semantic segmentation [48], panoptic segmentation [41], and image captioning [19].

We propose to extend the use of CLIP to the field of structural and civil engineering. By harnessing the transferable feature representations from CLIP, our approach achieves robust and accurate building attribute extraction across various residential areas and regions.

Zero-shot learning in visual tasks. Most deep learning models are confined to concepts learned during training. Zero-shot learning [40] aims to discern novel concepts from data never seen during training. [46] introduces a framework for learning image and text embeddings in a joint space. [22] converts a text embedding into semantic features which are taken as guidance for segmentation. [14, 20, 24] apply a generative model to produce the semantic features of unseen classes.

We tap into the broad visual and text knowledge CLIP and SAM have to robustly extract building attribute descriptions from images without any further training.

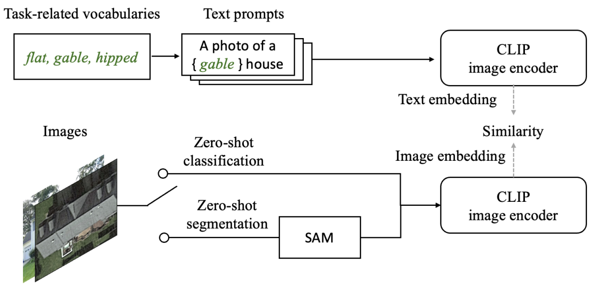


Figure 4. Our workflow utilizes the generalized image captioning of CLIP for zero-shot classification and zero-shot segmentation. The CLIP predicts the vocabulary by measuring cosine similarity between the textual embedding and visual embedding.

3. Zero-Shot Building Attribute Extraction

Our zero-shot workflow for building attribute extraction (Fig. 3) uses pre-trained large-scale vision and language models, CLIP and SAM, to achieve robust and versatile image understanding performance without needing any human annotations. It has two components: image-level captioning developed for image classification, and segment-level captioning for image segmentation.

3.1. Large-scale Models: CLIP and SAM

As a large-scale vision and language model pre-trained on extensive sets of image-text pairs, CLIP [31] comprises an image encoder denoted as Ψ and a text encoder denoted as Φ. These components are trained together to map input images and texts into a common representation space. The training objective of CLIP is centered on contrastive learning, taking the correct image-text pairs as positive samples while treating the mismatched pairs as negative samples.

SAM [23] is a large-scale vision model pre-trained with more than 1 billion masks on 11 million images for image segmentation. This pre-training provides SAM with a deep understanding of various objects and scenes and strong zero-shot generalization. SAM supports multiple types of prompts such as points, boxes, and texts, and is capable of segmenting any object in an image given with certain prompts.

Given an image or an image region segmented by SAM, our method queries CLIP with a list of captions that pertain to a task in the fields of structural and civil engineering and then selects the most matching caption. This strategy allows us to extract various building attributes with a single model and strong robustness to image variations, whereas BRAILS would need one model for one task and suffer a performance drop when test images do not look like training images.

3.2. Our Zero-shot Workflow

Our initial step involves obtaining the relevant vocabularies based on user-provided requests. Suppose we possess a collection of tasks pertinent to structural and civil engineering areas. Users can choose specific tasks, such as *the roof shape of the building* or *the semantic parsing of the facade*. Upon receiving the selected tasks, we evaluate their suitability for employment in the zero-shot classification task or the zero-shot segmentation task. Following this, we apply the proposed zero-shot workflow to derive the building attributes associated with these chosen tasks.

Zero-shot classification. Given a street view image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, we input it into a pre-trained image encoder Ψ to derive its image embedding $\mathbf{E} = \Psi(\mathbf{I}) \in \mathbb{R}^{1 \times L}$, where L is the size of the image embedding. Let $\mathcal{T}$ represent a task from the users. We also have a group of task-specific categories $\{\mathcal{V}_1, \cdots, \mathcal{V}_{N_t}\}$ relevant to structural and civil engineering, where $\mathcal{V}_i$ represents the textual vocabulary of the i-th category and N_t is the number of categories in the task $\mathcal{T}$. Each of these categories is integrated into the prompt and sent into a pre-trained text encoder Φ to acquire the corresponding text embedding $\mathbf{F}^{V_i} = \Phi(\mathcal{V}_i) \in \mathbb{R}^{1 \times L}$. Both the image encoder and text encoder are derived from the CLIP model. Then we obtain a score vector $\mathcal{S}_i$ by computing the similarity between the image embedding $\mathbf{E}$ and the text embedding $\mathbf{F}^{V_i}$ indicated by

$$\mathcal{S}_i = \texttt{sim}(\mathbf{E}, \mathbf{F}^{V_i}), \tag{1}$$

where $\texttt{sim}(\mathbf{E}, \mathbf{F}^{V_i})$ represents the cosine similarity, computed by the dot product between l_2-normalized $\mathbf{E}$ and $\mathbf{F}^{V_i}$, i.e., $\texttt{sim}(\mathbf{E}, \mathbf{F}^{V_i}) = \mathbf{E}^\top \mathbf{F}^{V_i} / \|\mathbf{E}\| \|\mathbf{F}^{V_i}\|$. On this basis, we identify the most suitable category index for $\mathbf{I}$, denoted by $\mathcal{P}_{img}(\mathbf{I})$, with the highest score in $\mathcal{S}$ by

$$\mathcal{P}_{img}(\mathbf{I}) = \arg\max_i \mathcal{S}_i. \tag{2}$$

Zero-shot segmentation. We take advantage of SAM for segmenting residential objects including *roofs*, *fence*, *doors*, *windows*, and *facades*. Let $\{\mathcal{C}_1, \cdots, \mathcal{C}_{N_c}\}$ denote the categories for the residential objects, where $\mathcal{C}_k$ represents the textual vocabulary of the k-th category and N_c is the number of categories. We take a street view image $\mathbf{I}$ as input to SAM, which in turn produces category-agnostic non-overlapped binary masks $\mathcal{M} = \{\mathbf{M}_j \mid \mathbf{M}_j \in \mathbb{B}^{H \times W}\}_{j=1}^{N}$, where N indicates the number of the masks. Next, we need to assign these binary masks with the corresponding semantic categories. Specifically, we design a zero-shot semantic segmentation workflow using the strong image captioning capacity of CLIP. Given a binary mask $\mathbf{M}_j$, a masked image $\mathbf{I}^{M_j}$ is obtained by element-wise multiplication presented by

$$\mathbf{I}^{M_j} = \mathbf{M}_j \odot \mathbf{I}. \tag{3}$$

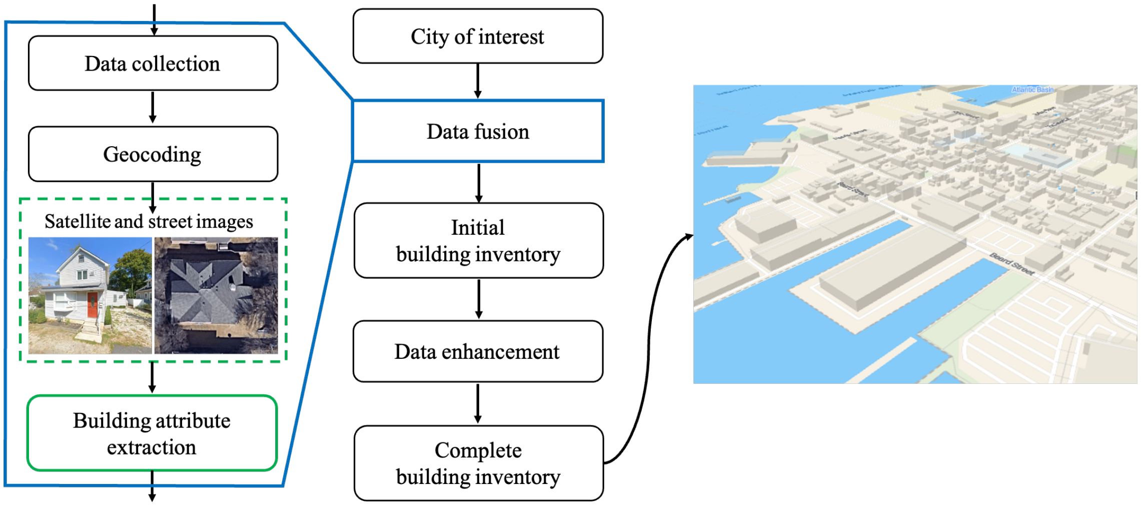

Data collection
Geocoding
Satellite and street images
Building attribute extraction
City of interest
Data fusion
Initial building inventory
Data enhancement
Complete building inventory

Figure 5. The framework pipeline of BRAILS offers a standardized approach to construct realistic databases of building inventories.

Subsequently, $\mathbf{I}^{M_j}$ is fed into the image encoder Ψ to produce the masked image embedding $\mathbf{E}^{M_j} = \Psi(\mathbf{I}^{M_j})$. Simultaneously, each of the categories is incorporated into the prompt and sent into the pre-trained text encoder Φ to get the embedding $\mathbf{F}^{C_k} = \Phi(\mathcal{C}_k)$. With the masked image embedding $\mathbf{E}^{M_j}$ and the category embedding $\mathbf{F}^{C_k}$, a similarity score $\mathcal{R}_j$ is computed by measuring the similarity between $\mathbf{E}^{M_j}$ and $\mathbf{F}^{C_k}$ using

$$\mathcal{R}_{j,k} = \texttt{sim}(\mathbf{E}^{M_j}, \mathbf{F}^{C_k}), \tag{4}$$

where $\texttt{sim}$ is defined in the same way as in Eq. 1. Our zero-shot segmentation outputs a segmentation map $\mathcal{P}_{seg}(\mathbf{I}) \in \mathbb{R}^{H \times W}$ computed by

$$\mathcal{P}_{seg}(\mathbf{I})^{(h,w)} = \arg\max_{k} \left\{ \mathcal{R}_{j,k} \mid \mathbf{M}_j^{(h,w)} = 1 \right\}, \tag{5}$$

where $\mathcal{P}_{seg}(\mathbf{I})^{(h,w)}$ is the predicted category index at the pixel position (h, w). Note that $\mathbf{M}_j^{(h,w)} = 1$ means that we take the binary mask $\mathbf{M}_j$ that shows 1 at the pixel position (h, w).

4. Experiments

Our work begins by explaining how BRAILS extracts building attributes from building inventory databases. We then present a variety of tasks, including determining the number of floors, classifying roof types, identifying the year of construction, and parsing facades. BRAILS employs a random split for its training and validation datasets, evaluating on the validation set. Therefore, it presents excellent performance on the validation set but shows decline on the novel domain. Our approach, on the other hand, uses a single module to extract building attributes for various tasks without the need for human-labeled data. We collect our own data as a novel domain and demonstrate that our method generalizes better to new domains.

4.1. BRAILS Framework

BRAILS [2, 7] offers a standardized approach to constructing realistic databases of building inventories, facilitating the creation of building information models at a regional scale. The framework of BRAILS is composed of multiple stages illustrated in Fig. 5.

Collecting regional-scale building information often requires the utilization of multiple resources. These resources encompass images, point clouds, property tax records, crowd-sourced maps, etc. These resources may be owned by different entities and stored in diverse formats. The data fusion process in Fig. 5 aims to create fused building information data that surpasses the original data in terms of informativeness and comprehensiveness.

The data collection module in the data fusion involves the integration of multiple building information datasets to yield information that is more consistent, precise, and practical than what any single source can provide. The outcome of data collection still lacks building attribute information due to data scarcity. For instance, crowd-sourced maps often lack completeness, particularly in rural areas, while they tend to be more comprehensive in densely urbanized regions and areas targeted for humanitarian mapping interventions. Similarly, administrative databases containing property tax assessment records often exhibit missing entries. This deficiency is a common issue across nearly all data sources. To handle these issues, the BRAILS framework applies the building attribute extraction module (Sec. 4.2) based on the initial database to predict the absent building attributes and effectively fill in the missing values.

The initial building inventory generated by the data fu-

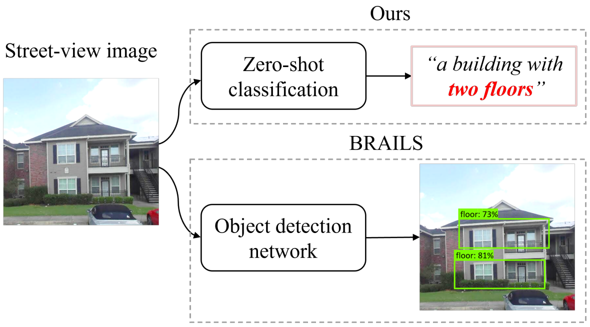


Figure 6. BRAILS requires a supervised object detection network to detect the number of floors, whereas our zero-shot workflow is capable of making direct inferences on the input images without any fine-tuning.

Table 1. BRAILS presents superior scores on the roof type classification task. Note that the BRAILS's scores are essentially training accuracy due to a lack of split over the training and the validation set. Our score is approaching BRAILS on the flat type.

Accuracy (%)	# Images	BRAILS	Ours	*Our Gain*
Gable type	8449	**99.2**	2.0	-97.2
Hip type	8451	**99.4**	47.8	-51.6
Flat type	8447	**99.6**	98.1	-1.5
Micro-Average	25347	**99.4**	49.2	-50.2
Macro-Average	25347	**99.4**	49.2	-50.2

sion process might still remain incomplete. For example, occlusions caused by trees or cars can impact predictions related to building attributes, such as the number of floors. This occlusion can subsequently lower the precision of predictions obtained from the building attribute extraction model. Therefore, a data enhancement module is implemented to address these incomplete predictions, aiming to make valuable contributions toward achieving a comprehensive final building inventory.

4.2. Building Attribute Extraction

We assume the preliminary collection of indexing information including building addresses, coordinates, year of construction, and structural style was established. With the indexing information in hand, it becomes feasible to utilize the Google Maps API for retrieving satellite and street-view images corresponding to each respective building.

The existing BRAILS framework comprises multiple modules to extract building attributes (absent in the original database) from these images. These modules leverage deep learning architectures and are designed for the image classification task and the semantic segmentation task, refined through supervised learning.

However, the BRAILS method for building attribute extraction has two main limitations. First, it requires human annotation for all tasks, and these annotations can be expensive and time-consuming. Second, the performance of the models in BRAILS is evaluated on validation datasets that are similar to the training set. However, these models' generalization is not carefully studied. The ability of these models to generalize to new and unseen data from different geographical locations depends on how similar the building inventory for these locations is to the data used to train the BRAILS models. Generalization is a complex topic that has been actively researched recently. In this work, we propose a new zero-shot workflow that addresses these challenges by using large-scale models that are trained with self-supervised techniques on a variety of datasets and can be adapted to a wide range of downstream tasks.

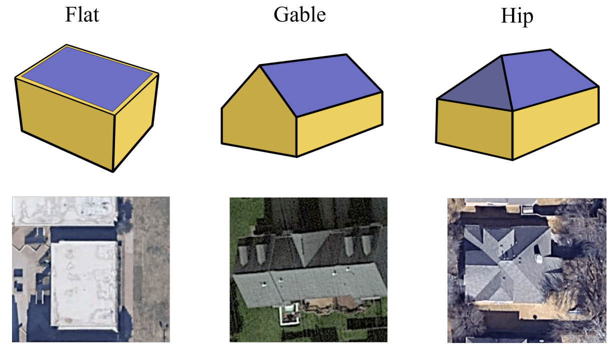


Figure 7. The common roof types and example satellite images for these types.

4.3. Roof Type Classification

Dataset. We present the three primary categories of roof types commonly used globally: flat, gabled, and hipped (Fig. 7). The whole dataset collected in BRAILS contains $6,000$ labeled satellite images with $2,000$ images for each of the three roof types. The dataset is randomly split into the training and validation set.

Baseline. To identify the roof types of every building in the region, BRAILS adopts an *image classification network* with a ResNet-50 [18] backbone. Different from the roof type classification module in BRAILS, our workflow treats this task as a *zero-shot classification* task and does not require pre-training on these training data.

Results. Table 1 shows results on different roof types: gable, hip, and flat types. Note that due to the lack of train-test split information, the results from BRAILS in Table 1 represent the training accuracy. On average for roof type classification, BRAILS presents 99.4% of accuracy whereas our method indicates a performance of 49.2%. Our score is approaching BRAILS on the flat type. There are two reasons for this performance gap. First, BRAILS utilizes a supervised image classification model and encounters a small domain gap between the training and validation set. In contrast, our method is a zero-shot workflow that

Table 2. The scores of BRAILS and ours for the year built classification are computed using BRAILS' own validation set. BRAILS presents superior performance because it adopts supervised training with numerous human-annotated data, whereas our method requires neither human annotations nor any fine-tuning.

Year range	# Images	BRAILS	Ours	*Our Gain*
Pre 1969	30198	**62.0**	38.7	-23.3
1970 - 1979	10485	**11.6**	0.8	-10.8
1980 - 1989	20519	10.8	**12.8**	+2.0
1990 - 1999	13537	8.3	**46.3**	+38.0
2000 - 2009	19178	**14.0**	0.1	-13.9
Post 2010	5944	**1.6**	0.0	-1.6
Micro-Average	99861	**26.1**	20.7	-5.4
Macro-Average	99861	**18.1**	16.4	-1.7

does not require any pre-training on annotated data. Second, our workflow utilizes the image captioning ability of CLIP which is less generalized to satellite images.

4.4. Year Built Classification

Dataset. The objective of the year built classification task involves categorizing buildings into various groups, each representing a distinct range of construction years. The dataset devised for year built classification consists of $56,660$ annotated street-view images and has a random split for training and validation set. We consider 6 distinct ranges of year built constructions: before 1969 (Pre 1970), 1970-1979, 1980-1989, 1990-1999, 2000-2009, and after 2010 (Post 2010). We compare ours against BRAILS on the BRAILS' own validation set for the year built classification task.

Baseline. In the BRAILS framework, this task is considered by using an *image classification network* with ResNet-50 [18] as the backbone. In contrast, our approach adopts a *zero-shot classification* method, omitting any need for supervised learning with the training data.

Results. In Table 2, on average for the year built classification, the BRAILS method presents an accuracy of 26.1%, while our method shows 20.7%. This performance gap reflects that the BRAILS is trained in such annotated data with a random training and testing split and is not affected by the domain gap between the training and validation set. *Our proposed zero-shot workflow requiring neither annotations nor training already gets close to BRAILS' supervised learning method on the year-built classification task.*

4.5. Facade Parsing

Dataset & Task. Building facade parsing needs to segment the building facade from a street view image into multiple semantic categories, such as the roof, windows, doors, and facades. This can be used to extract more detailed building attributes and contribute to a complete building inventory. The existing BRAILS method for facade parsing requires pixel-wise annotations from humans, which can be costly and time-consuming.

Table 3. Our method is compared with existing open-vocabulary segmentation methods OVSeg and ODISE on facade parsing task. The mIoU (%) performance is evaluated on the validation set of BRAILS.

mIoU (%)	Roof	Door	Window	Facade	*Mean*
OVSeg [25]	50.9	66.4	79.1	38.8	57.3
ODISE [42]	53.1	64.1	83.0	44.7	60.2
Ours	**55.6**	**67.9**	**85.5**	**48.6**	**61.5**

Table 4. The scores of BRAILS and ours for detecting the number of floors are evaluated on BRAILS' own validation set. BRAILS shows superior performance because it adopts supervised training and encounters few domain gaps with the validation set, whereas our method requires neither human annotation nor additional fine-tuning.

Accuracy (%)	# Images	BRAILS	Ours	*Our Gain*
One-story	2393	**88.5**	80.8	-7.7
Two-story	580	56.4	**57.8**	+1.4
Three-story	16	**56.3**	0.0	-56.3
Micro-Average	2989	**82.0**	75.9	- 6.1
Macro-Average	2989	**67.0**	46.2	-20.8

Baseline. We provide an ablation study of our zero-shot segmentation on the dataset in BRAILS for the facade parsing task. We choose recent state-of-the-art open-vocabulary semantic segmentation methods OVSeg [25] and ODISE [42] as our baseline models. OVSeg requires large amounts of annotated data and requires fine-tuning over CLIP model, and ODISE involves a training guided by the text-to-image diffusion UNet [33].

Results. Table 3 shows the comparison results of ours with the baseline models OVSeg and ODISE. Our method presents a mIoU (mean intersection over union) of 61.5 which is the highest among all the baseline methods that require annotated data or extra guidance. This is due to the strong image segment capability from SAM model in our zero-shot segmentation. *It indicates that we apply zero-shot segmentation to achieve facade parsing without requiring annotated data or additional fine-tuning. Moreover, it allows us to segment images that have never been seen before, even if the building images are from different styles or regions.*

4.6. # Floors

Dataset. The dataset collected in BRAILS comprises $60,000$ street view images sourced from various counties in New Jersey in the United States, excluding Atlantic County. These data are then partitioned into three subsets: a training set, a validation set, and a testing set, randomly distributed in proportions of 80%, 15%, and 5% of the total data, respectively.

Baseline. Within the BRAILS framework, an *object de-*

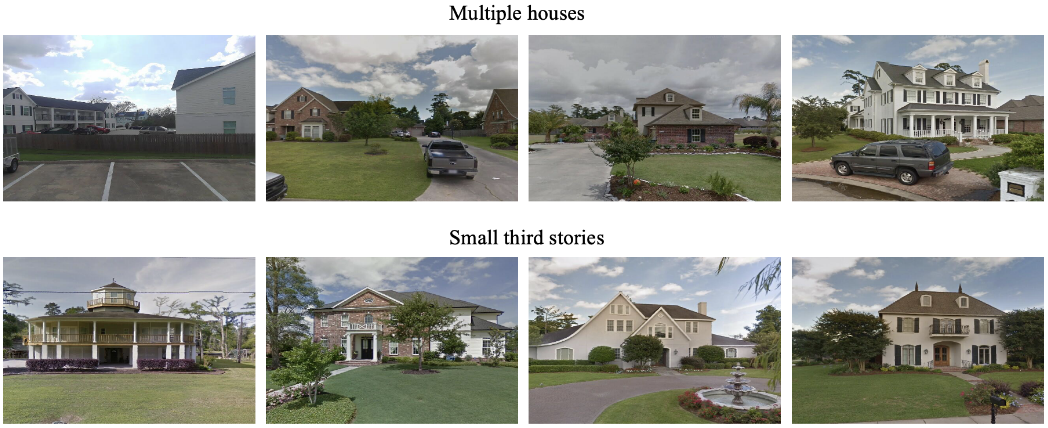


Figure 8. The exemplary images of the three-story houses from BRAILS' validation set where our method completely failed.

Table 5. BRAILS suffers from a clear performance drop due to the domain gap from the novel domain, while our method presents a robust performance on the novel domain data for detecting the number of floors.

Accuracy (%)	# Images	BRAILS	Ours	*Our Gain*
One-story	210	70.7	**77.6**	+6.9
Two-story	198	55.2	**74.0**	+18.8
Three-story	37	33.3	**50.0**	+16.7
Micro-Average	445	66.5	**76.1**	+9.6
Macro-Average	445	53.07	**67.2**	+14.13

tection network based on EfficientDet-D4 architecture [36] is employed to identify visible floors in the street-view images. In contrast, our workflow tackles this task as a *zero-shot classification* problem and we eliminate the need for any preliminary training on the training set as shown in Fig. 6.

Results. According to Table 4, on BRAILS' own validation set, the average accuracy of BRAILS is 82.0% while the accuracy of our method is 75.9%. The difference in accuracy between the two methods is due to the fact that the BRAILS method is a supervised object detection model trained on multiple data with annotated bounding boxes. In contrast, our method is a zero-shot workflow that does not require any training or human annotations. We completely failed on three-story houses in Table 4. The reasons are threefold: 1) there are only a few images for the three-story house; 2) some images contain multiple houses; 3) the third stories are usually very small. We present the exemplary images for these failing cases in Fig. 8.

Generalization. We further evaluate the generalization of the BRAILS method and ours over novel domains. As BRAILS takes data that mostly comes from the cities on the West Coast of the U.S. for training, it is reasonable that BRAILS shows good performance in these regions. We postulate that there exists domain gap between West Coast and East Coast images. Therefore, we select some image data from the cities on the East Coast as a novel domain. Specifically, we collect street view images from Houston, Texas and randomly select 460 images as the novel domain for testing. The results of the BRAILS method and our method on the novel domain are shown in Table 5. Our method presents an average accuracy of 76.1% and BRAILS shows 66.5% on the data from the novel domain. This result confirms that BRAILS suffers from a generalization gap due to regional variations in building appearances. *Our proposed zero-shot workflow provides a more robust, promising baseline to novel domains without model supervision or prompt tuning.*

5. Conclusion

In this paper, we proposed a new zero-shot workflow for building attribute extraction in the structural and civil engineering domains. Our workflow utilizes large-scale vision and language models, CLIP and SAM, to generate captions for buildings from satellite and street-view images. This allows us to extract building attributes without relying on human annotations. Our workflow has several advantages over existing methods. First, it is scalable and robust to regional variations in visual and geometrical appearances. Second, it can generalize to novel buildings in unseen regions. We evaluated our workflow on the datasets of building images with the task of zero-shot image classification and zero-shot segmentation. Our results demonstrate the effectiveness of our approach for building attribute extraction. In the future, we plan to take a more advanced way of utilizing large-scale models and turn our zero-shot workflow into a specific expert model in structural and civil engineering areas.

Acknowledgements. This work is funded in part by NSF 2131111: Natural Hazards Engineering Research Infrastructure: Computational Modeling and Simulation Center.

References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 3

[2] Barbaros Cetiner, Charles Wang, Frank McKenna, Sascha Hornauer, and Yuhui. Guo. Nheri-simcenter/brails: Version v3.0.0. zenodo, 2023. 5

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 3

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 3

[6] GG Deierlein and Á Zsarnóczay. State-of-art in computational simulation for natural hazards engineering. zenodo, 2019. 1, 2

[7] Gregory G Deierlein, Frank McKenna, Adam Zsarnóczay, Tracy Kijewski-Correa, Ahsan Kareem, Wael Elhaddad, Laura Lowes, Matthew J Schoettler, and Sanjay Govindjee. A cloud-enabled application framework for simulating regional-scale impacts of natural hazards on the built environment. *Frontiers in Built Environment*, 6:558706, 2020. 5

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 3

[9] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015. 3

[10] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in Neural Information Processing Systems*, 27, 2014. 3

[11] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113, 2017. 2

[12] Lluis Gomez, Yash Patel, Marçal Rusinol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4230–4239, 2017. 3

[13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 3

[14] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. Context-aware feature generation for zero-shot semantic segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1921–1929, 2020. 3

[15] Yunhui Guo, Chaofeng Wang, Stella X Yu, Frank McKenna, and Kincho H Law. Adaln: A vision transformer for multidomain learning and predisaster building information extraction from images. *Journal of Computing in Civil Engineering*, 36(5):04022024, 2022. 1, 2

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6, 7

[19] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 3

[20] Ping Hu, Stan Sclaroff, and Kate Saenko. Uncertainty-aware learning for zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 33:21713–21724, 2020. 3

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[22] Naoki Kato, Toshihiko Yamasaki, and Kiyoharu Aizawa. Zero-shot semantic segmentation via variational mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 4

[24] Peike Li, Yunchao Wei, and Yi Yang. Consistent structural relation learning for zero-shot segmentation. *Advances in Neural Information Processing Systems*, 33:10317–10327, 2020. 3

[25] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 7

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 3

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 3

[28] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 3

[29] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[30] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021. 3

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015. 3

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 7

[34] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 153–170. Springer, 2020. 3

[35] Albert Gordo Soldevila and Diane Larlus-Larrondo. Leveraging captions to learn a global visual representation for semantic retrieval, Dec. 27 2018. US Patent App. 15/633,892. 3

[36] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10781–10790, 2020. 8

[37] Rebekka Volk, Julian Stengel, and Frank Schultmann. Building information modeling (bim) for existing buildings—literature review and future needs. *Automation in Construction*, 38:109–127, 2014. 2

[38] Chaofeng Wang, Sascha Hornauer, Stella X Yu, Frank McKenna, and Kincho H Law. Instance segmentation of soft-story buildings from street-view images with semiautomatic annotation. *Earthquake Engineering & Structural Dynamics*, 52(8):2520–2532, 2023. 1, 2

[39] Chaofeng Wang, Qian Yu, Kincho H Law, Frank McKenna, X Yu Stella, Ertugrul Taciroglu, Adam Zsarnóczay, Wael Elhaddad, and Barbaros Cetiner. Machine learning-based regional scale intelligent modeling of building information for natural hazard risk management. *Automation in Construction*, 122:103474, 2021. 2, 3

[40] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2018. 3

[41] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 3

[42] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 7

[43] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 27, 2014. 3

[44] Qian Yu, Chaofeng Wang, Barbaros Cetiner, Stella X Yu, Frank Mckenna, Ertugrul Taciroglu, and Kincho H Law. Building information modeling and classification by visual learning at a city scale. *arXiv preprint arXiv:1910.06391*, 2019. 1, 2

[45] Qian Yu, Chaofeng Wang, Frank McKenna, Stella X Yu, Ertugrul Taciroglu, Barbaros Cetiner, and Kincho H Law. Rapid visual screening of soft-story buildings from street view images using deep learning classification. *Earthquake Engineering and Engineering Vibration*, 19:827–838, 2020. 1, 2

[46] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2002–2010, 2017. 3

[47] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017. 3

[48] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 3