MMLANDMARKS: a Cross-View Instance-Level Benchmark for Geo-Spatial Understanding

Supplementary Material

We provide additional information and visualizations for the MMLANDMARKS dataset. Sec. A describes the distribution of images per landmark, geographic and categoric distributions, and details the collection pipeline for selecting the landmarks with OpenStreetMaps, Wikimedia Commons and Wikipedia. In Sec. B, more visualizations of the dataset are included, with bounding box examples and a more detailed analysis of each modality illustrated with landmarks from Denver in Figures 10 and 11. The VLM data processing procedure is mentioned in Sec. C, with examples of when the VLM fails to correctly categorise the images in Fig. 9. Finally, additional illustrations of landmarks from MMLANDMARKS are presented in Figures 12-15.

Algorithm 1 MMLANDMARKS landmark collection pipeline

Require: OSM data information, Wikidata, Wikimedia Commons, Wikipedia

$Dataset = \{\}$
for each sample **polygon**$_i$ in OSM **do**
  (lat,lon) = nodes(**polygon**$_i$)
  if (`wikidata` $\cup$ `wikipedia`) $\in$ **polygon**$_i$ tags **then**
    Extract Wiki information Q_i.
    (Commons$_i$ & Wikipedia$_i$) $\leftarrow$ `wikidata.org/wiki/Qi`
    if $\exists$(ground images) $\in$ Commons$_i$ **then**
      $G_i = True$
    end if
    if $\exists$(wiki text) $\in$ Wikipedia$_i$ **then**
      $T_i = True$
    end if
    if abs$(min(lat) - max(lat)) \cap$ abs$(min(lon) - max(lon)) < 400m$ **then**
      $S_i = True$
      $C_i = True$
    end if
  end if
  MML$_i = \{G_i, T_i, S_i, C_i\}$
  if $\{M_i = True | \forall M_i \in$ MML$_i\}$ **then**
    $Dataset \leftarrow$ MML$_i$
  end if
end for
$Dataset$ contains US landmarks with ground, satellite, text, and GPS information.

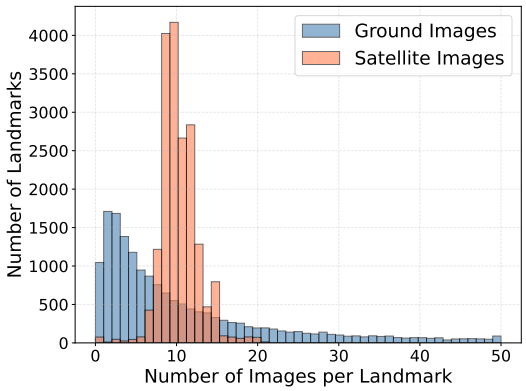


Figure 4. Histogram distribution of the number of images per landmark. A large proportion of landmarks have between 1 and 10 ground images, with a long-tailed distribution. The number of satellite images per landmark follows a bell curve centred at 10 images, with some landmarks having up to 20 aerial images.

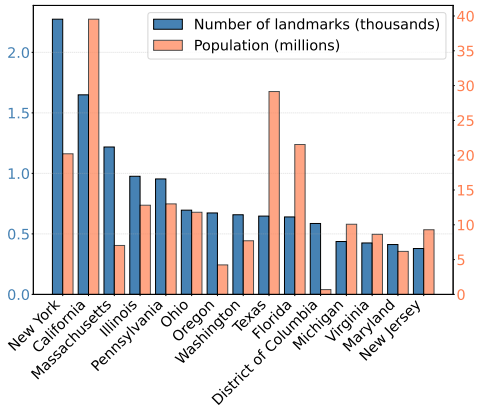
A. Dataset Details

Distribution. As mentioned in Sec. 3, the data contains long-tail distributions, both in terms of geographical location, landmark type, and number of images per landmark. In this section, we provide additional information and illustrate these characteristics of the datasets that are inherent in geo-spatial data.

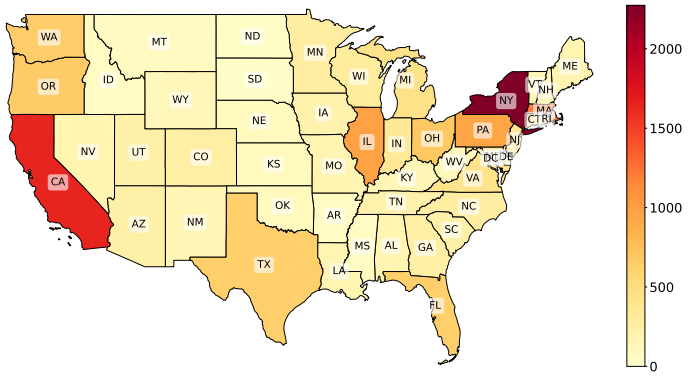
In Fig. 4, the number of images per landmark is plotted for ground and satellite images. The ground images have a long tail, with most landmarks having 1 to 10 images. The satellite images have a normal distribution, with an average of 10 images per landmark, and some have as many as 20 images, spanning over two decades.

The histogram in Fig. 5 shows the top 15 states with the most landmarks in blue, as well as the population of each state in orange. The number of landmarks does not correlate with the population of the state, with a clear example of the District of Columbia (a federal district which is not part of any state), where many governmental landmarks are located despite a low number of residents. The states with the lowest number of landmarks are Hawaii (22), South Dakota (35), and North Dakota (37).

On the right of Fig. 5, a map of the United States illustrates the geographic distribution of landmarks by state, showing the highest concentrations in states with larger populations, such as New York, California, Texas, and Florida.



(a) Histogram distribution of the top 15 states with the most landmarks (blue), compared to each state's population (orange). The two distributions diverge, with certain states being disproportionately over-represented or under-represented.



(b) Geographic visualization of the number of landmarks per state in the United States. similar to (a), a large proportion of landmarks come from popular areas such as New York & California.

Figure 5. Visual and Geographical illustrations of the landmark distribution across MMLANDMARKS.

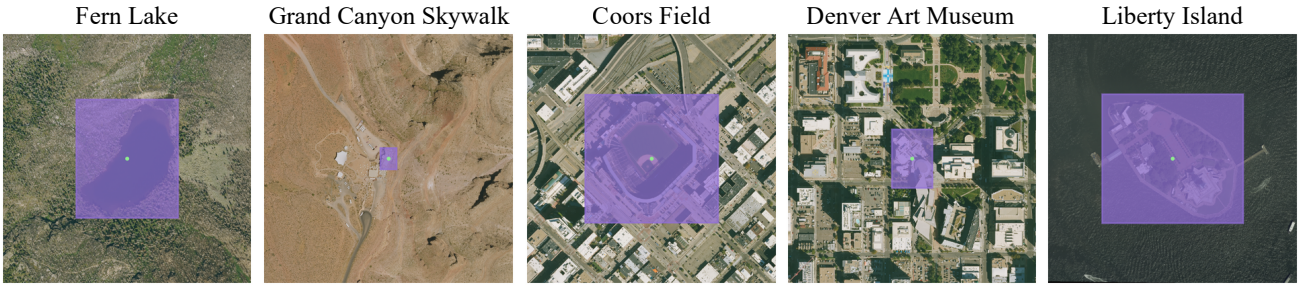


Figure 6. Visualization of the center GPS (green) and bounding boxes (purple) for the polygons associated with different landmarks.

Collection pipeline. In Alg. 1, we present our landmark collection pipeline for MMLANDMARKS. For each polygon found from OpenStreetMaps (OSM), we filter those that have either a `wikidata` or `wikipedia` tag in their description. This grants us the possibility to search for the landmark's Wiki-identifier (Q_i = Q123456), from which the landmark's Wikipedia and Wikimedia Commons webpage can be retrieved.

If ground images and Wikipedia text are available, we proceed to the final part of the processing pipeline: aerial uniformity. We aim to streamline the collection process from the National Agricultural Imagery Program (NAIP) while ensuring that all images have consistent dimensions. In fact, many landmarks vary in size and shape, sometimes spanning several kilometers (e.g., lakes, railways, mountain ranges, etc.). The scarcity of such large landmarks would cause geospatial ambiguity: to capture the entire landmark, one may need to sample images from a small zoom level, leading to low-resolution imagery. Alternatively, samples from various locations of the landmark at high resolution could be a direction. However, this approach often results in a predominance of uniform images depicting only rocks, water, or small portions of a building.

This both complicates the pipeline and incurs ambiguity due to the lack of correspondence between the modalities. To this end, we heuristically choose a maximum length of 400 meters for the largest side of the landmark's bounding

Licenses	Counts
CC BY	70,288
CC SA	85
CC BY-SA	181,409
CC BY-NC-SA	76
Public Domain	56,216
CC0	16,046
Attribution	873
No restrictions	3,902

Table 7. Specific licenses for the ground view images in MMLANDMARKS. The licenses and link to each image are also included as part of the dataset for proper attribution.

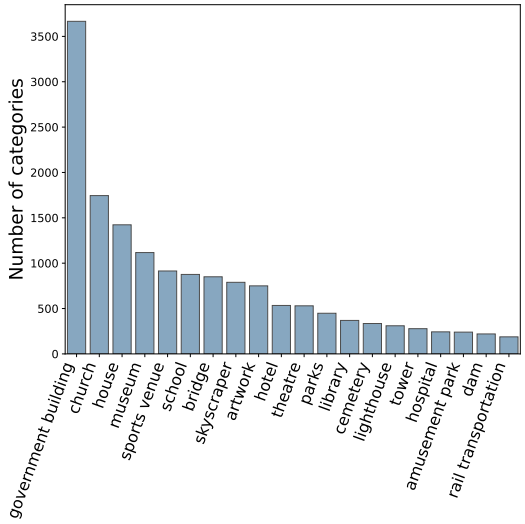


Figure 7. 15 most popular super-categories in MMLANDMARKS. Each landmark is categorized as one of the 79 hierarchical super-categories from Ramzi et al. [68]. There is a large majority of "government building" landmarks, which are further analysed in Fig. 8.

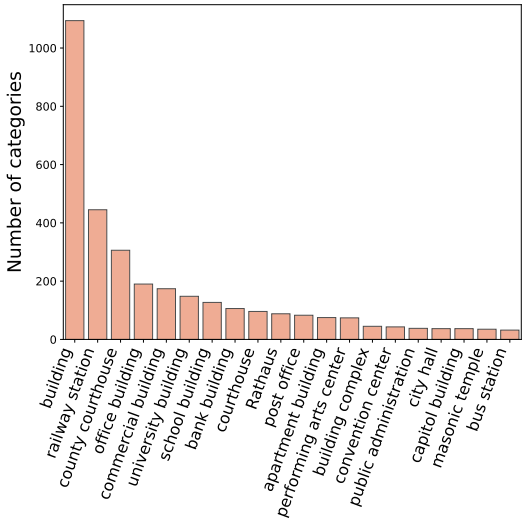


Figure 8. 15 most popular categories from the "government building" super-category. Since a CLIP text encoder is used to classify each category amongst one of the 79 hierarchical categories, all landmark categories with the tag "building" are placed under "government building".

box. As mentioned in Sec. 3, the ground images are resized such that the largest side is 800 pixels. Keeping the same pixel dimensions, we aim to sample aerial images of pixel size 800×800. After visual inspections, we find that 400 meters yields a balance between the landmark and its surroundings.

Licenses. We collect the MMLANDMARKS dataset, intending to make everything freely accessible and usable. The NAIP aerial imagery is categorized as Public Domain, making it ideal for data collection and incurring no significant filtration in the process. For ground-view images sourced from Wikimedia Commons, an additional step is required to verify the licensing of each image. We keep all the images licensed under "Creative Commons", "Public Domain", and other licenses such as "Attribution" and "No restrictions" which allow copy, modification, and redistribution. We group the licenses and summarize their count in Tab. 7.

B. Additional Visualizations

Super-categories. As extra information retrieved from each landmark's Wikidata page, we collect tags under the "Instance of" section for all landmarks as their category, similar to the Hierarchical GLDv2 [68], in which 79 unique hierarchical super-categories are defined. Once the tags are found, we use a CLIP text encoder to embed each category and assign it to one of the super-categories from Ramzi et al. [68]. This is achieved by computing the logits between the 79 super-category embeddings and each query category, then selecting the super-category with the highest probability. In Fig. 7, we show the number of landmarks per super-category. Because of the large proportion of landmarks identified as "government building", we also illustrate the sub-categories of the top super-category (see Fig. 8). We can see that since a lot of landmarks are simply labelled as "building", when grouped, they are placed under the "government building" label.

More examples. We show the diversity in MMLANDMARKS by providing examples of landmarks in Fig. 10, as well as temporal changes in the collected aerial imagery in Fig. 11. We illustrate the intra-state diversity by showing landmarks located around Denver, Colorado.

Figure 10 illustrates the diversity of images associated with each landmark, a characteristic typical of web-sourced imagery. For example, in the case of "Coors Field", ground-level photos include views captured before the game (image 2), during the game (images 3–7), and after the game (image 1), exhibiting variations in camera angle and lighting conditions. The "Denver Art Museum" ground views also include scanned documents (image 8) and art from the museum (image 4). Even more intriguing are images 2 and 3 from the Red Rocks Amphitheatre, where the second image is a postcard version from the landmark, from the same angle as the third image. Finally, the last landmark, "Denver Union Station", reflects the diverse angles and locations from which images are taken of the same landmark, which together give a full understanding of the landmark and its surroundings. Fig. 10 also illustrates the geo-spatial fine

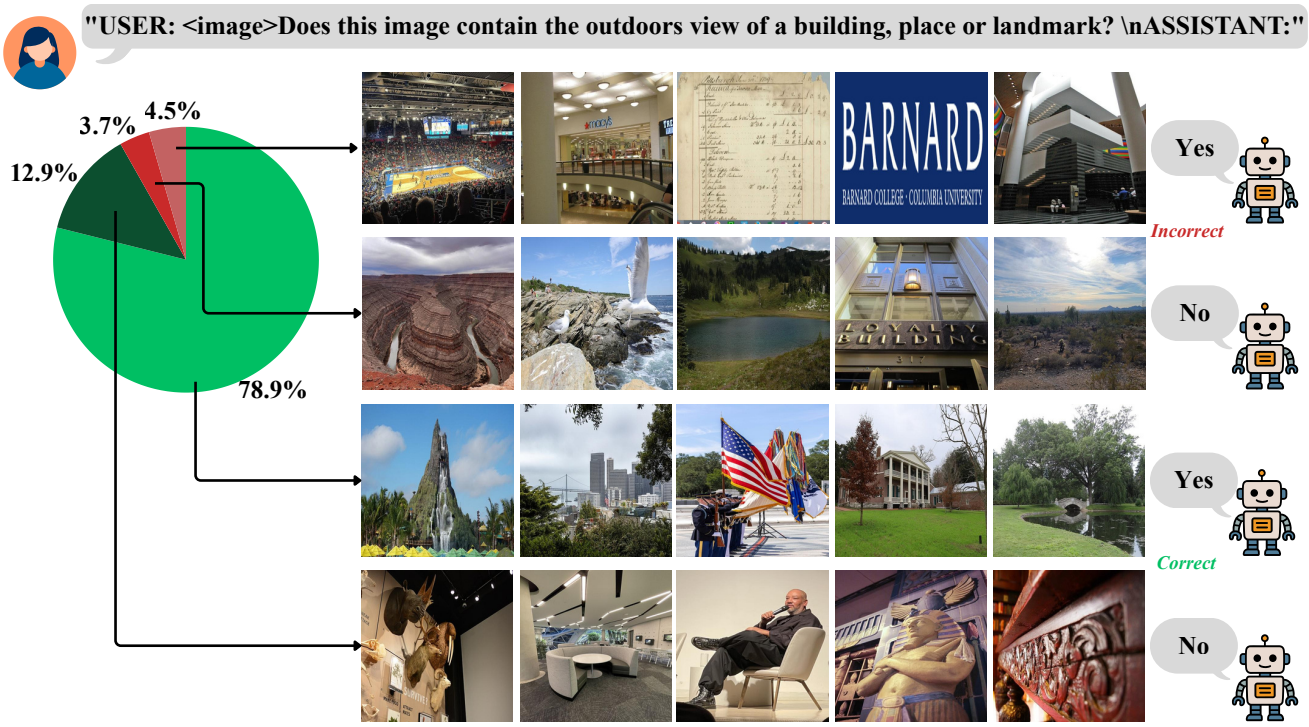


Figure 9. VLM Filtering: Examples of wrong categorizations during the VLM - indoor/outdoor filtering stage. The model may incorrectly classify images as belonging to the opposite category due to visual ambiguities in the images. Five examples are shown from the 8.2% wrongly classified ground images in the 1000 visually inspected samples.

granularity by providing accurate GPS locations, along with long textual descriptions of the landmark.

Similarly, in Fig. 11, the aerial images taken from the NAIP illustrate the diversity of the same location when inspecting the same place through time. Not only are all images different in terms of capture time, angle, sun orientation, and season, but physical changes also happen, which can be detected by comparing the images together. Most obvious is the "Fern Lake", where the size and color of the water vary. It is also evident that tree density was significantly higher in the older (top) images compared to the newer (bottom) views. The fourth image seems to show traces of the wildfires, which have burnt down a majority of the surrounding trees.

In columns one and four, clear changes appear in the surroundings, with a building appearing in the top left corner of the Coors Field images between images 2 and 3, or a parking space being removed between images 4 and 5.

For the "Red Rocks Amphitheatre", an asphalt parking lot was added between the times when images 1 and 2 were captured. Additionally, the top of the Amphitheatre is now covered by a greyish roof, which is also visible in ground-level images found through a web search for the landmark. The MMLANDMARKS is extremely diverse in terms of geography, textual, visual, and temporal complexity. We show additional examples of our dataset in Figures 12-15. MMLANDMARKS offers the possibility to train and evaluate models in a unified framework that more realistically reflects our world.

C. VLM Data Processing

In Fig. 9, we show the splitting process used to create the subset with only outdoor images (83% of the ground images classified as outdoors, as mentioned in Sec. 3. We use a Vision Language Model (`llava-hf/llava-1.5-7b-hf`) [46] and prompt the following: "USER: <image> Does this image contain the outdoors view of a building, place, or landmark?\n ASSISTANT:". The VLM's answering window is made very small, such that the VLM only replies by "Yes" (outdoors) or "No" (indoors). In doing so, we remove all images where there are no obvious cues in the image that depict the landmark seen from the exterior. After all images are processed, we sample 1000 ground views randomly and manually inspect them for soundness. We find that 8.2% are wrongly classified, with 3.7% wrongly categorised as indoors, and 4.5% wrongly categorised as outdoors.

However, the 3.7% that are wrongly classified as indoors and should have been marked as outdoors are actually relevant to filter out, since they are often close-up, natural im-

ages that do not provide any particular information about the landmark, and would therefore not contribute during training had they been correctly labelled. This is reflected in the bottom row of Fig. 9, where the VLM classifies a close view of a building or natural landscapes where it cannot see any noticeable landmark, as "No" landmark.

For the images from an indoor setting that are incorrectly classified as "outdoors", many wrong conclusions from the model stem from basketball courts, malls, or the interior of large buildings that share similar architecture as outdoor landmarks (respectively seen in the top row of Fig. 9: images one, two, and five). Some scanned documents and paintings of buildings are also falsely classified as outdoor images. Upon inspection, we conclude that the number of wrong classifications is acceptably low. The results for models trained on the outdoors subset and presented in Tab. 6 also demonstrate the relevance of employing such a filtration process as a pre-processing clean-up of the raw MM-LANDMARKS dataset.

Figure 10. Additional examples of landmarks from MMLANDMARKS. The four landmarks are sampled from around Denver, Colorado.

Figure 11. Visual diversity in the aerial imagery from landmarks in MMLANDMARKS. The images are arranged chronologically, ranging from older images at the top to newer images at the bottom. More images of the landmarks are included in the dataset, which are not presented here. The landmarks are sampled from around Denver, Colorado. The temporal aspect of the dataset augments the instance with different moments of capture, angle, and weather conditions. Zoom in to see fine-grained changes in the urban surroundings of columns 1, 2, and 5.

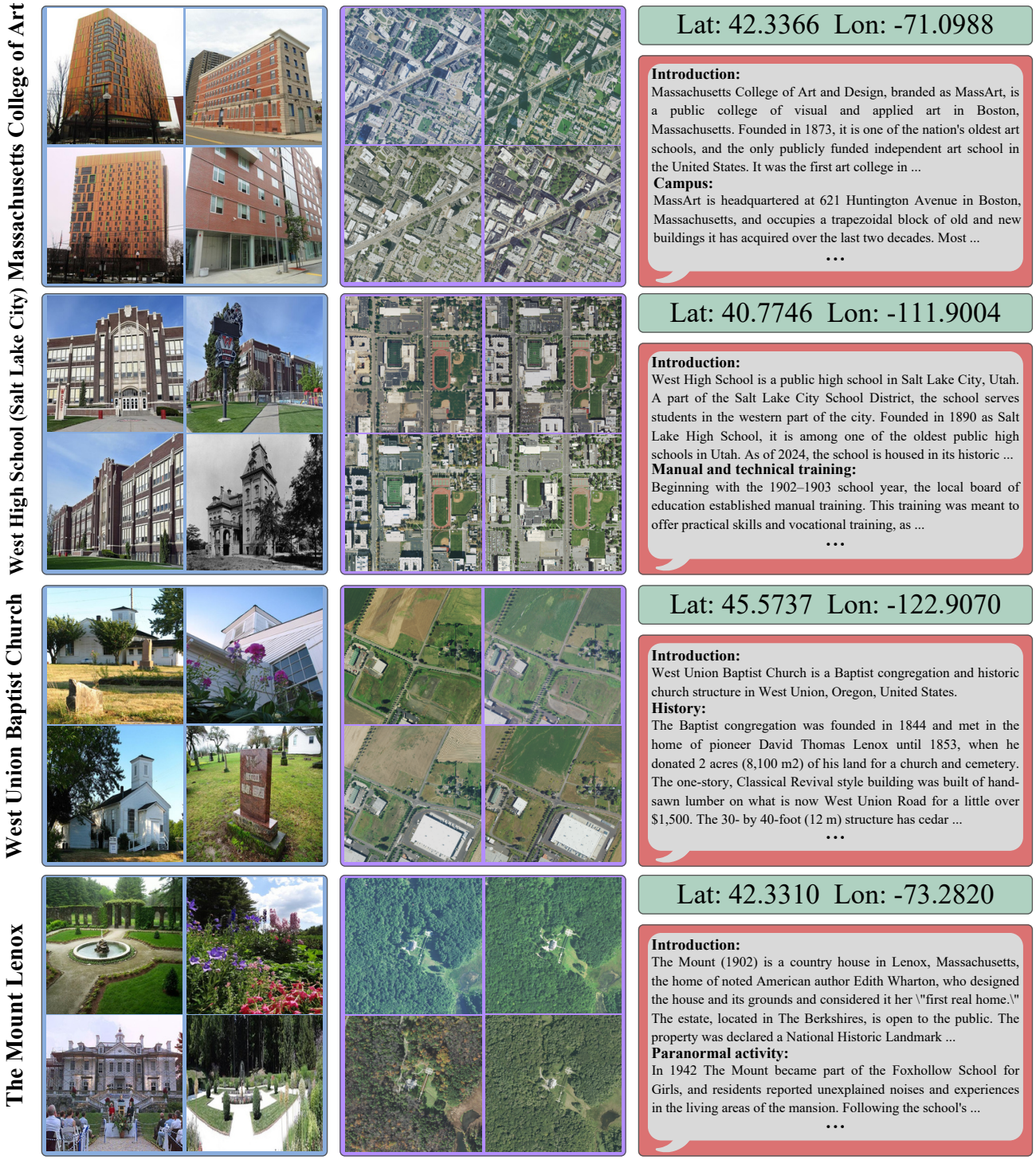


Figure 12. Additional examples from the MMLANDMARKS dataset. We illustrate the diversity in the dataset by randomly sampling landmarks. We show sample ground and satellite views, as well as the exact GPS location and parts of the textual descriptions.

Figure 13. Additional examples from the MMLANDMARKS dataset. We illustrate the diversity in the dataset by randomly sampling landmarks. We show sample ground and satellite views, as well as the exact GPS location and parts of the textual descriptions.

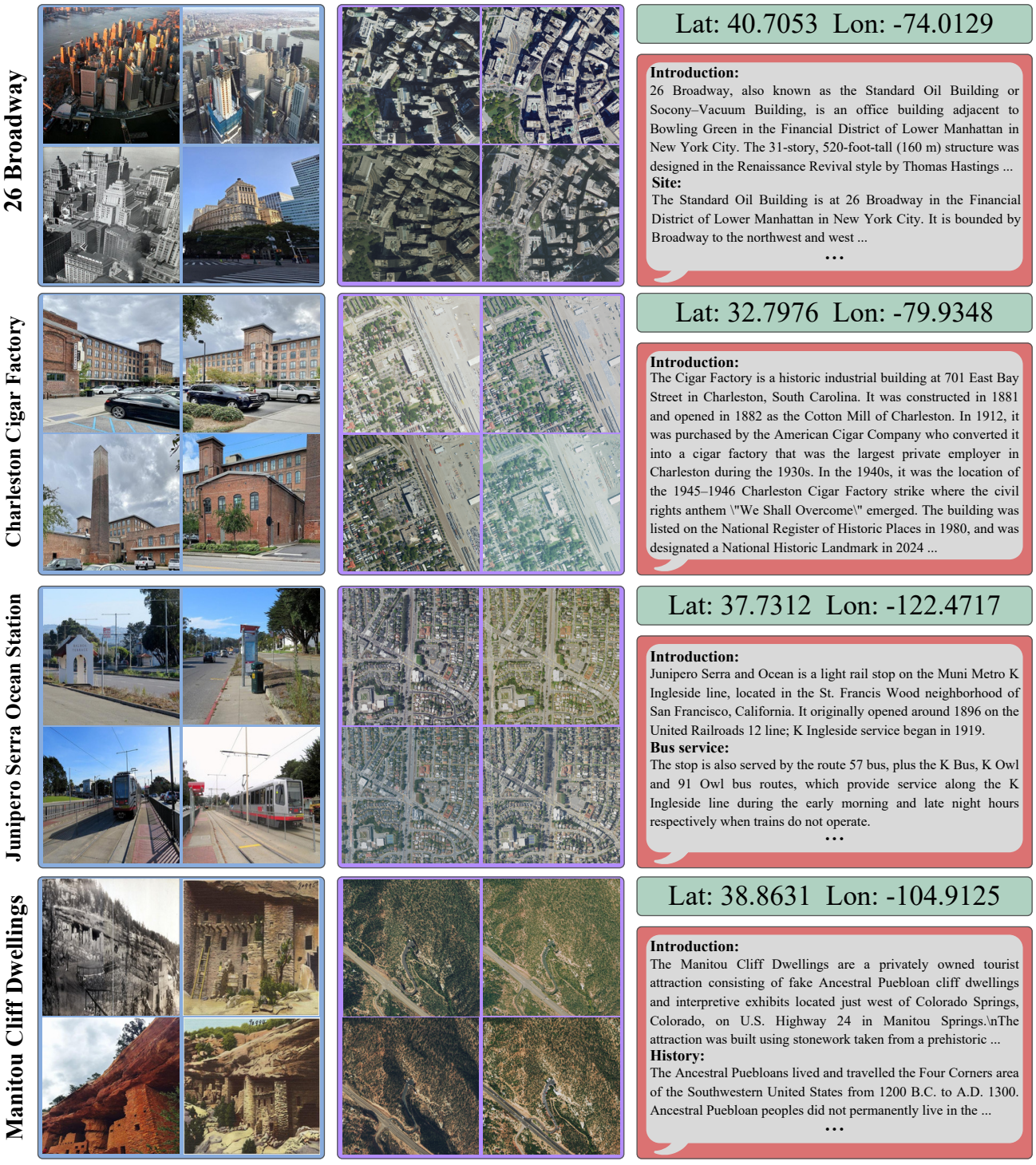


Figure 14. Additional examples from the MMLANDMARKS dataset. We illustrate the diversity in the dataset by randomly sampling landmarks. We show sample ground and satellite views, as well as the exact GPS location and parts of the textual descriptions.

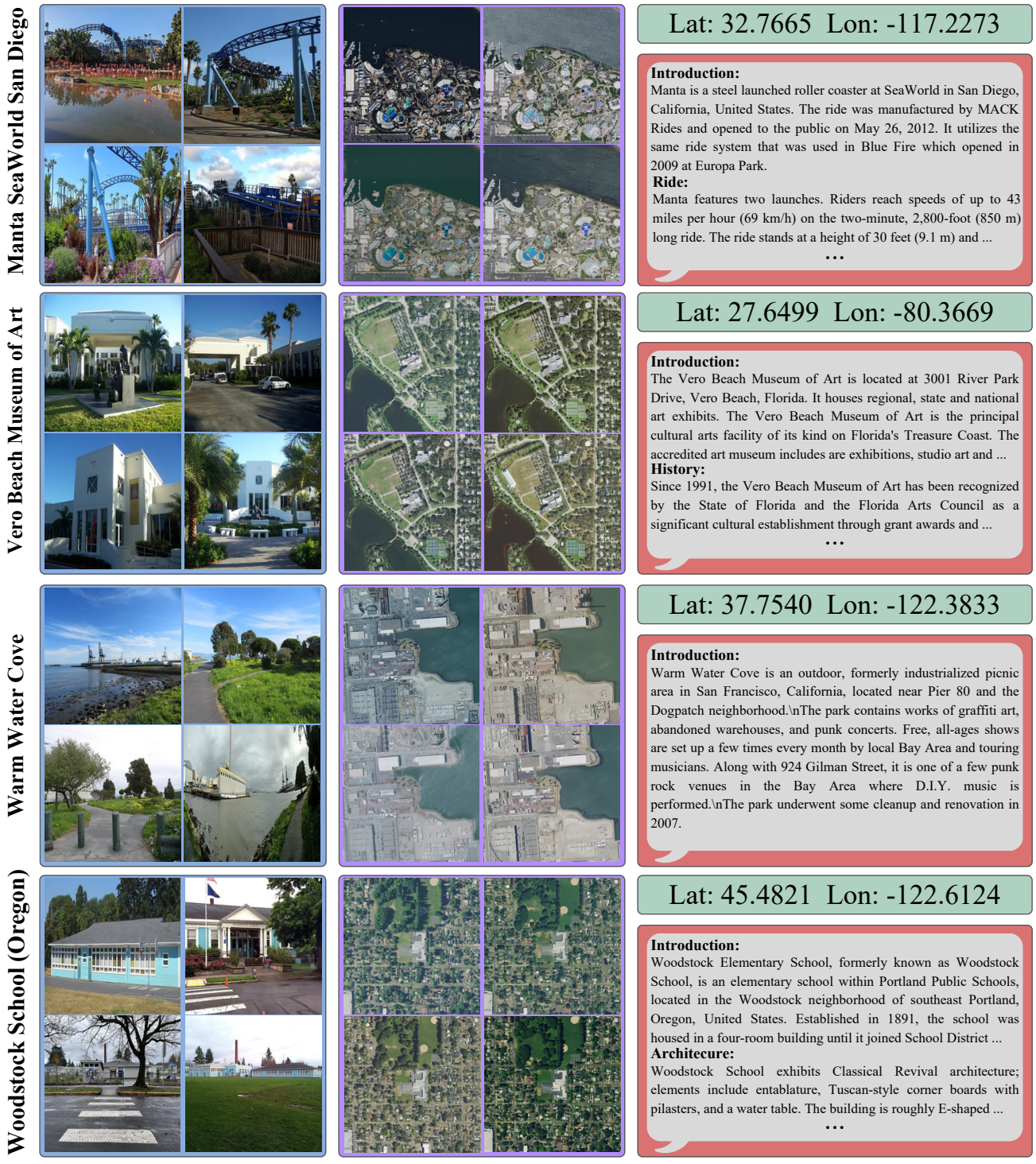


Figure 15. Additional examples from the MMLANDMARKS dataset. We illustrate the diversity in the dataset by randomly sampling landmarks. We show sample ground and satellite views, as well as the exact GPS location and parts of the textual descriptions.