

# 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

- 1:  $Dataset = \{\}$
- 2: **for** each sample  $polygon_i$  in *OSM* **do**
- 3:    $(lat, lon) = nodes(polygon_i)$
- 4:   **if**  $(wikidata \cup wikipedia) \in polygon_i$  tags **then**
- 5:     Extract Wiki information  $Q_i$ .
- 6:      $(Commons_i \& Wikipedia_i) \leftarrow wikidata.org/wiki/Q_i$
- 7:     **if**  $\exists(\text{ground images}) \in Commons_i$  **then**
- 8:        $G_i = True$
- 9:     **end if**
- 10:    **if**  $\exists(\text{wiki text}) \in Wikipedia_i$  **then**
- 11:      $T_i = True$
- 12:    **end if**
- 13:    **if**  $abs(min(lat) - max(lat)) \cap abs(min(lon) - max(lon)) < 400m$  **then**
- 14:      $S_i = True$
- 15:      $C_i = True$
- 16:    **end if**
- 17:    **end if**
- 18:     $MML_i = \{G_i, T_i, S_i, C_i\}$
- 19:    **if**  $\{M_i = True | \forall M_i \in MML_i\}$  **then**
- 20:      $Dataset \leftarrow MML_i$
- 21:    **end if**
- 22: **end for**
- 23: *Dataset* contains US landmarks with ground, satellite, text, and GPS information.

---

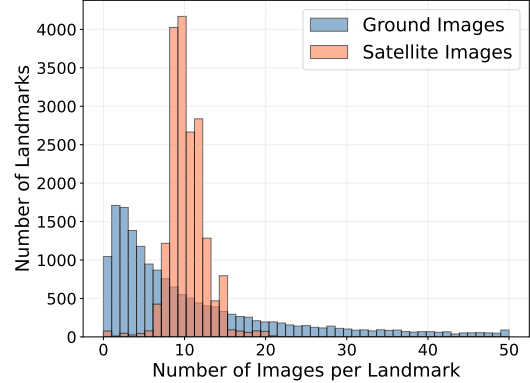


Figure 4. Histogram distribution of the number of images per landmark. A large proportion of landmarks have between 1 and 10 ground images, with a long-tailed distribution. The number of satellite images per landmark follows a bell curve centred at 10 images, with some landmarks having up to 20 aerial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

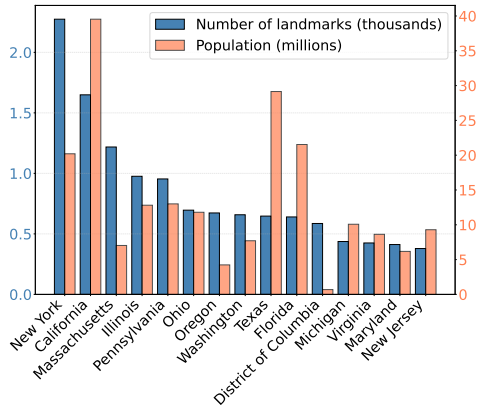
### A. Dataset Details

**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.

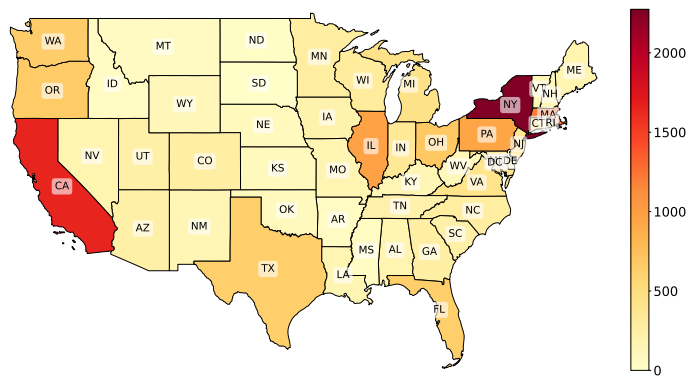
In Fig. 4, the number of images per landmark is plotted for ground and satellite images. The ground images have a long tail, with most landmarks having 1 to 10 images. The satellite images have a normal distribution, with an average of 10 images per landmark, and some have as many as 20 images, spanning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

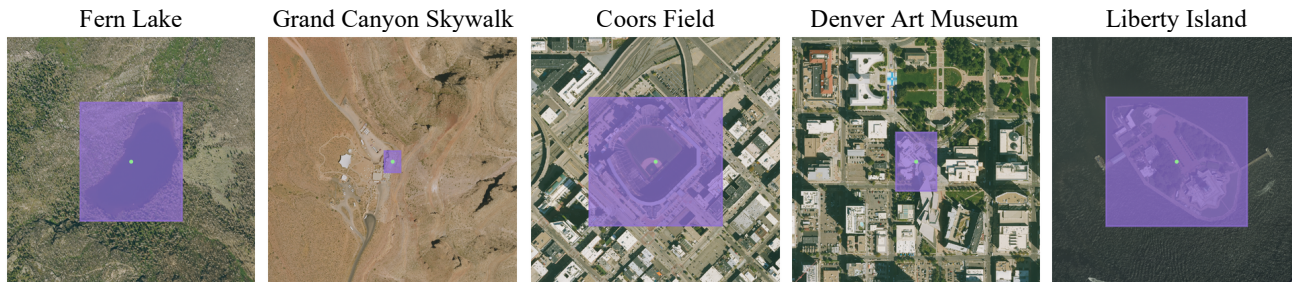


Figure 6. Visualization of the center GPS (green) and bounding boxes (purple) for the polygons associated with different landmarks.

**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

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.

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

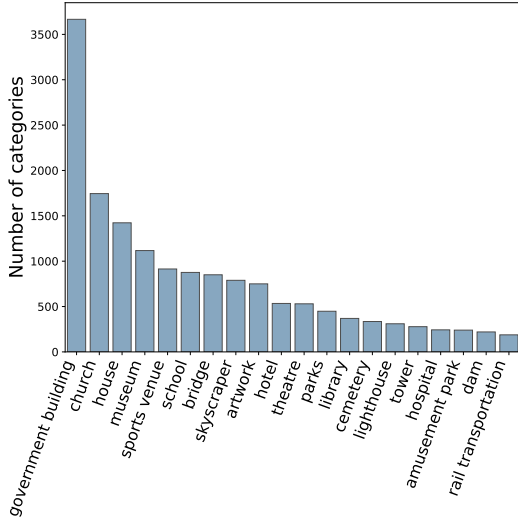


Figure 7. 15 most popular super-categories in MMLANDMARKS. Each landmark is categorized as one of the 79 hierarchical super-categories from Ramzi et al. [68]. There is a large majority of “government building” landmarks, which are further analysed in Fig. 8.

box. As mentioned in Sec. 3, the ground images are re-sized such that the largest side is 800 pixels. Keeping the same pixel dimensions, we aim to sample aerial images of pixel size  $800 \times 800$ . 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,

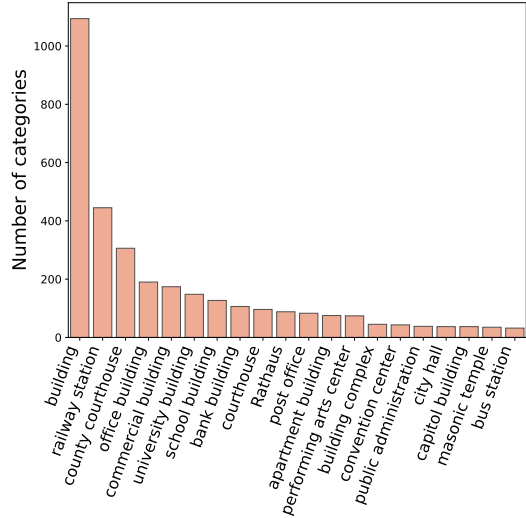


Figure 8. 15 most popular categories from the “government building” super-category. Since a CLIP text encoder is used to classify each category amongst one of the 79 hierarchical categories, all landmark categories with the tag “building” are placed under “government building”.

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

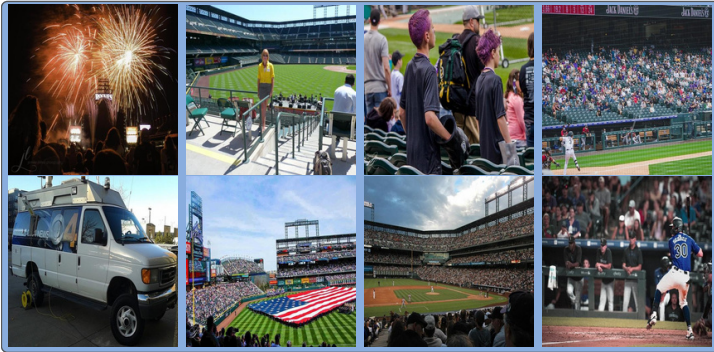
In Fig. 9, we show the splitting process used to create the subset with only outdoor images (83% of the ground images classified as outdoors, as mentioned in Sec. 3. We use a Vision Language Model (llava-hf/llava-1.5-7b-hf) [46] and prompt the following: “USER: <image> Does this image contain the outdoors 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.

Coors Field



Lat: 39.7560 Lon: -104.9940

**Introduction:**  
Coors Field is a baseball stadium in downtown Denver, Colorado, United States. It is the ballpark of Major League Baseball's Colorado Rockies. Opened in 1995, the park is located in Denver's Lower Downtown neighborhood, two blocks from Union Station. The stadium has a ...

**Construction:**  
Coors Field was the first new stadium added in a six-year period in which Denver's sports venues were upgraded, along with Ball Arena ...

...

Denver Art Museum



Lat: 39.7373 Lon: -104.9896

**Introduction:**  
The Denver Art Museum (DAM) is an art museum located in the Civic Center of Denver, Colorado. With an encyclopedic collection of more than 70,000 diverse works from across the centuries and world, the DAM is one of the largest art museums between the West Coast and Chicago. It is known for its ...

**1893–1923:**  
The museum's origins can be traced back to the founding of the Denver Artists Club in 1893. The Club renamed itself the Denver Art Association in ...

...

Red Rocks Amphitheatre



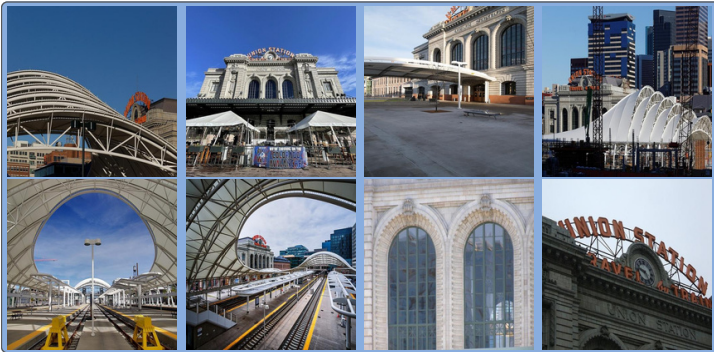
Lat: 39.6654 Lon: -105.2057

**Introduction:**  
Red Rocks Amphitheatre (also known colloquially as simply Red Rocks) is an open-air amphitheater in the western United States near Morrison, Colorado, approximately ten miles (16 km) southwest of Denver. It is owned and operated by the city of Denver. In addition to several other large sandstone ...

**History:**  
The natural features surrounding the amphitheater were formed millions of years ago as part of the Fountain Formation, then lifted and tilted during a geological upheaval ...

...

Denver Union Station



Lat: 39.7532 Lon: -105.0003

**Introduction:**  
Denver Union Station is the main railway station and central transportation hub in Denver, Colorado. It is located at 17th and Wynkoop Streets in the present-day LoDo district and includes the historic station house, a modern open-air train shed, a 22-gate underground bus station, and light rail station. A station ...

**19th century: Original structures:**  
Denver's first train station was constructed in 1868 to serve the new Denver Pacific Railway, which connected Denver to the main transcontinental line at Cheyenne, Wyoming. By 1875 ...

...

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.

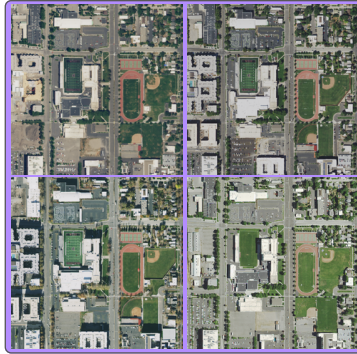
West High School (Salt Lake City) Massachusetts College of Art



Lat: 42.3366 Lon: -71.0988

**Introduction:**  
 Massachusetts College of Art and Design, branded as MassArt, is a public college of visual and applied art in Boston, Massachusetts. Founded in 1873, it is one of the nation's oldest art schools, and the only publicly funded independent art school in the United States. It was the first art college in ...  
**Campus:**  
 MassArt is headquartered at 621 Huntington Avenue in Boston, Massachusetts, and occupies a trapezoidal block of old and new buildings it has acquired over the last two decades. Most ...  
 ...

West Union Baptist Church



Lat: 40.7746 Lon: -111.9004

**Introduction:**  
 West High School is a public high school in Salt Lake City, Utah. A part of the Salt Lake City School District, the school serves students in the western part of the city. Founded in 1890 as Salt Lake High School, it is among one of the oldest public high schools in Utah. As of 2024, the school is housed in its historic ...  
**Manual and technical training:**  
 Beginning with the 1902–1903 school year, the local board of education established manual training. This training was meant to offer practical skills and vocational training, as ...  
 ...

The Mount Lenox



Lat: 45.5737 Lon: -122.9070

**Introduction:**  
 West Union Baptist Church is a Baptist congregation and historic church structure in West Union, Oregon, United States.  
**History:**  
 The Baptist congregation was founded in 1844 and met in the home of pioneer David Thomas Lenox until 1853, when he donated 2 acres (8,100 m2) of his land for a church and cemetery. The one-story, Classical Revival style building was built of hand-sawn lumber on what is now West Union Road for a little over \$1,500. The 30- by 40-foot (12 m) structure has cedar ...  
 ...

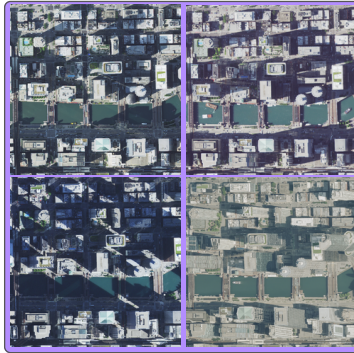


Lat: 42.3310 Lon: -73.2820

**Introduction:**  
 The Mount (1902) is a country house in Lenox, Massachusetts, the home of noted American author Edith Wharton, who designed the house and its grounds and considered it her "first real home."  
 The estate, located in The Berkshires, is open to the public. The property was declared a National Historic Landmark ...  
**Paranormal activity:**  
 In 1942 The Mount became part of the Foxhollow School for Girls, and residents reported unexplained noises and experiences in the living areas of the mansion. Following the school's ...  
 ...

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.

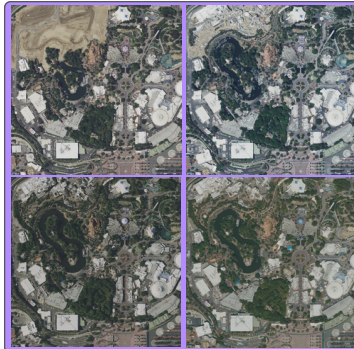
321 North Clark



Lat: 41.8883 Lon: -87.6306

**Introduction:**  
 321 North Clark at Riverfront Plaza is a 35-story, 155.45 m (510.0 ft) skyscraper constructed from 1983 to 1987 in Chicago, Illinois, United States. The tower was built by BCE Development Properties and designed by Skidmore, Owings & Merrill as part of the Riverfront Plaza development on the north bank of the Chicago River. 321 North Clark opened in April 1987 and was named "city development of the year" by the Chicago Sun-Times. The building was originally named Quaker Tower after its anchor tenant, the Quaker Oats Company ...

Golden Horseshoe Saloon

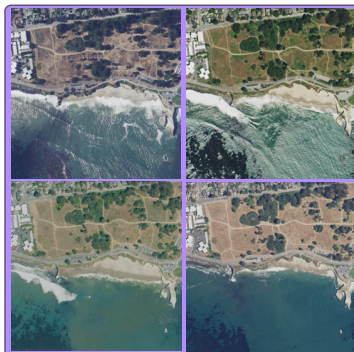


Lat: 33.8119 Lon: -117.9202

**Introduction:**  
 The Golden Horseshoe Saloon (referred to as Pecos Bill's Golden Horseshoe Saloon during construction) is a restaurant and attraction at Disneyland Park in Anaheim, California in the United States. It opened in 1955 with several other original attractions at Disneyland Park. Over the years the venue has housed ...

**History:**  
 The Golden Horseshoe Stage unofficially opened on July 13, 1955, as the Golden Horseshoe Saloon, when Walt and Lillian Disney, along with dozens of guests, celebrated ...

Lighthouse Field State Beach



Lat: 36.9518 Lon: -122.0298

**Introduction:**  
 Lighthouse Field State Beach is a protected beach in the state park system of California, United States. It is located in the city of Santa Cruz at the north end of Monterey Bay. The beach overlooks the Steamer Lane surfing hotspot. It also contains the Santa Cruz Surfing Museum, housed in a 1967 ...

**Natural history:**  
 Lighthouse Field State Beach is a wintering ground for migrating monarch butterflies. Other resident animals include California sea lions and American black swifts.

Malibu Hindu Temple



Lat: 34.0951 Lon: -118.7099

**Introduction:**  
 Malibu Hindu Temple is a Hindu temple in located in the Santa Monica Mountains, in the city of Calabasas near Malibu, California. Built in 1981 and dedicated to the Hindu deity Venkateswara, it features traditional South Indian style and serves as a centre for Hindu worship and cultural events in Southern ...

**Architecture and Deities:**  
 The complex has two temples – the upper temple with Venkateswara as the presiding deity and the lower temple with Shiva as the presiding deity. Both temples are constructed ...

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.

Manta SeaWorld San Diego



Lat: 32.7665 Lon: -117.2273

**Introduction:**  
Manta is a steel launched roller coaster at SeaWorld in San Diego, California, United States. The ride was manufactured by MACK Rides and opened to the public on May 26, 2012. It utilizes the same ride system that was used in Blue Fire which opened in 2009 at Europa Park.

**Ride:**  
Manta features two launches. Riders reach speeds of up to 43 miles per hour (69 km/h) on the two-minute, 2,800-foot (850 m) long ride. The ride stands at a height of 30 feet (9.1 m) and ...

...

Vero Beach Museum of Art



Lat: 27.6499 Lon: -80.3669

**Introduction:**  
The Vero Beach Museum of Art is located at 3001 River Park Drive, Vero Beach, Florida. It houses regional, state and national art exhibits. The Vero Beach Museum of Art is the principal cultural arts facility of its kind on Florida's Treasure Coast. The accredited art museum includes are exhibitions, studio art and ...

**History:**  
Since 1991, the Vero Beach Museum of Art has been recognized by the State of Florida and the Florida Arts Council as a significant cultural establishment through grant awards and ...

...

Warm Water Cove



Lat: 37.7540 Lon: -122.3833

**Introduction:**  
Warm Water Cove is an outdoor, formerly industrialized picnic area in San Francisco, California, located near Pier 80 and the Dogpatch neighborhood. The park contains works of graffiti art, abandoned warehouses, and punk concerts. Free, all-ages shows are set up a few times every month by local Bay Area and touring musicians. Along with 924 Gilman Street, it is one of a few punk rock venues in the Bay Area where D.I.Y. music is performed. The park underwent some cleanup and renovation in 2007.

Woodstock School (Oregon)



Lat: 45.4821 Lon: -122.6124

**Introduction:**  
Woodstock Elementary School, formerly known as Woodstock School, is an elementary school within Portland Public Schools, located in the Woodstock neighborhood of southeast Portland, Oregon, United States. Established in 1891, the school was housed in a four-room building until it joined School District ...

**Architecture:**  
Woodstock School exhibits Classical Revival architecture; elements include entablature, Tuscan-style corner boards with pilasters, and a water table. The building is roughly E-shaped ...

...

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.