Supplementary Materials: PSM: Learning Probabilistic Embeddings for Multi-scale Zero-shot Soundscape Mapping

Anonymous Authors

1 DATASET CREATION

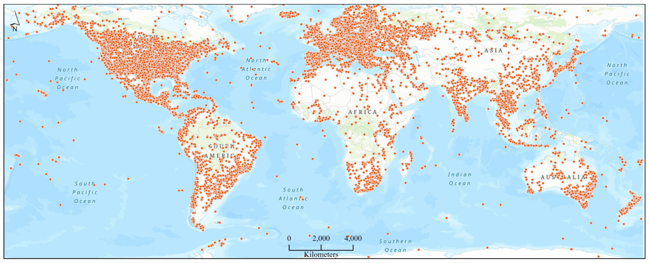

Figure 1: Distribution of samples in the *GeoSound* dataset.

We have created a new large-scale dataset (*GeoSound*) suitable for the task of zero-shot soundscape mapping, effectively increasing the size of available dataset [6] by more than 6-fold. To achieve this, we collected geotagged audios along with associated metadata (textual description, geolocation, time) from four different audio sources: *iNaturalist* [3], *YFCC100M* [8], *Radio Aporee* [4], and *Freesound* [2]. For each of the audio samples in our dataset, we downloaded 1500 × 1500 high-resolution (0.6m GSD) imagery from *Bing* and 1280 × 1280 low-resolution (10m GSD) *Sentinel-2 Cloudless* imagery from *EOX::Maps* [1]. Figure 1 illustrates the geospatial distribution of data samples in the *GeoSound* dataset worldwide.

1.1 Audio Sources

iNaturalist: This is an open-source platform for the community of Naturalists who upload observations for various species with records containing images, audio, and textual descriptions. We select observations with the flags: `Verifiable`, `Research Grade`, and `Has Sounds` to maximize data quality and completeness. This provides us with over 450k geotagged audios. To create a relatively balanced dataset with audio from different crowd-sourced platforms, we first only retain the species with at least 100 samples in our dataset. Then, we conduct round-robin random sampling of the observations, starting from the species with the lowest count and iteratively increasing the sample size until we reach our desired number of samples: 120k from 611 species. Finally, after a quality control filtering procedure, *iNaturalist* contributes 114 603 audios.

YFCC100M: YFCC100M is a publicly available, large multimedia dataset containing over 99 million images and around 0.8 million videos. This data is collected from the crowd-sourced platform *Flickr*. However, among the 0.8 million videos, only around 100k videos are found to be geotagged. Therefore, in our dataset, we extract audio from these geotagged videos only, contributing an additional 96 452 audio samples.

Radio Aporee: In our dataset, we also include the geotagged audios from the *SoundingEarth* dataset [6], which was built from the crowd-sourced platform hosted by the project *Radio Aporee::Maps*. This dataset contains field recordings of different types of audio from urban, rural, and natural environments. The *SoundingEarth* dataset contributes 49 284 audio samples.

Freesound: This is another commonly used platform for crowd-sourced audio containing field recordings from diverse acoustic environments. *Freesound* contributes a total of 48 680 audio samples.

split	iNaturalist	yfcc	aporee	freesound	total
train	108 753	92 055	46 893	46 318	294 019
val	1 851	1 565	797	787	5 000
test	3 999	2 832	1 594	1 575	10 000
total	114 603	96 452	49 284	48 680	309 019

Table 1: Distribution of *GeoSound* Dataset Across Splits and Audio Sources.

1.2 Data Split Strategy

We split our dataset to mitigate potential data leakage between data with similar locations in the training and validation/test sets. The distribution of data across training/validation/test sets and audio sources is given in Table 1. Our data split strategy on *GeoSound* dataset is described as follows:

1. We divide the world into 1° × 1° non-overlapping cells. This corresponds to the cell size of about 111km × 111km.
2. We only select the cells with at least 25 observations. The cells that do not pass this threshold are saved to be included in the train split of our dataset.
3. Based on the number of observations in each cell selected in step 2, we categorize them into three data density categories: *high*, *medium*, and *low* based on the 0.33 quantile and 0.66 quantile of the overall sample count from step 2.
4. For each category obtained in step 3, we randomly select 10% of cells to be held out for validation and test splits.
5. From the held-out cells obtained in step 4, we randomly sample 40% into validation split and the rest into test split.
6. For the validation/test split, 5000/10000 samples are randomly selected, matching the audio-source distribution of the train split.

2 UNCERTAINTY ESTIMATES

One of the advantages of PSM is that uncertainty estimates are automatically provided with representations of samples. After a sample is encoded, the σ associated with the distribution predicted by our framework represents its inherent uncertainty for any audio or satellite imagery. In Figure 2, we present sets of samples with high uncertainty and low uncertainty for *Bing* satellite imagery in our test set.

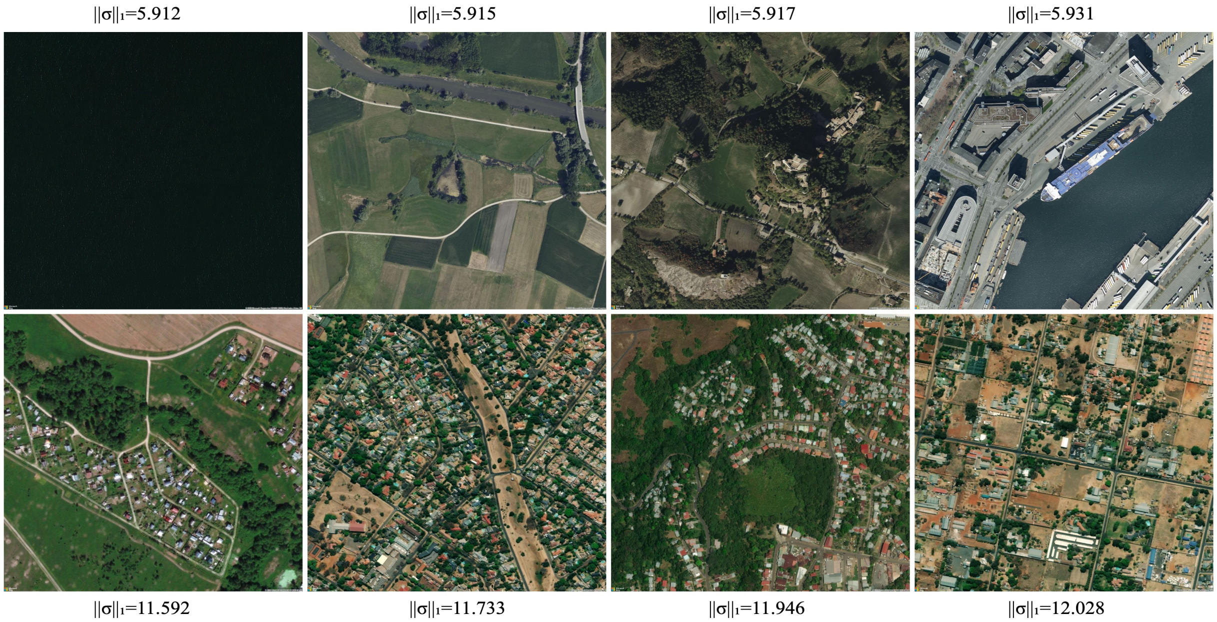

Figure 2: Uncertainty estimates are reflected by the $||\sigma||_1$ of selected samples from our *Bing* satellite imagery test set. These estimates are obtained from embeddings generated by our best-performing model trained on *Bing* imagery, without any additional metadata.

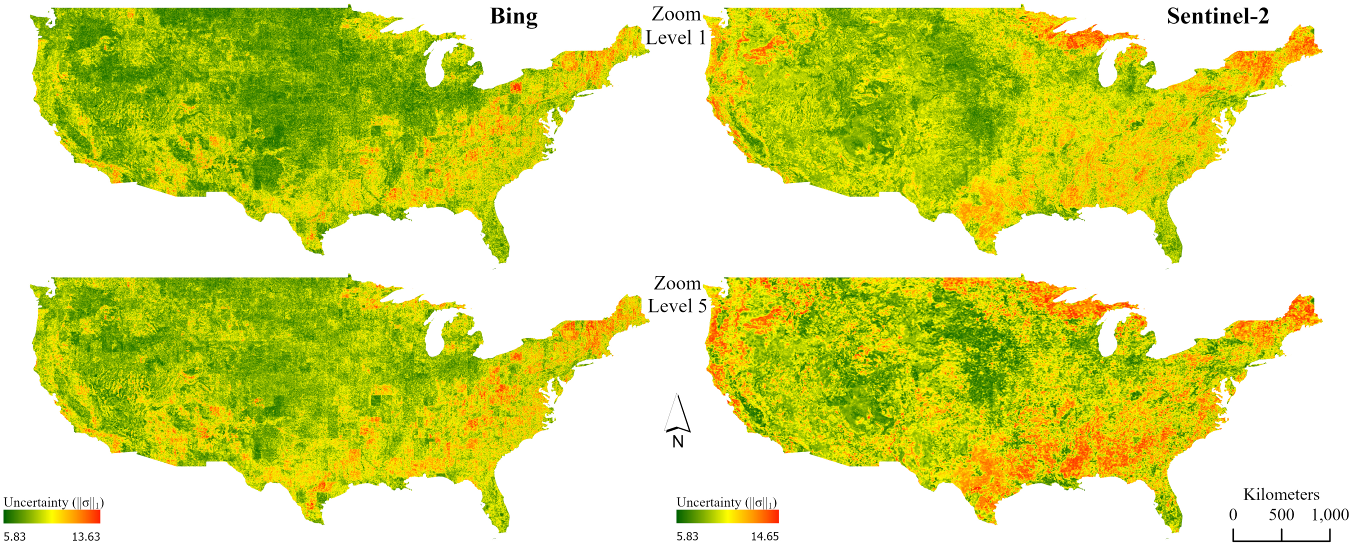

Figure 3: Uncertainty map of the satellite image embeddings for the USA. Uncertainty at each location is approximated as the $||\sigma||_1$ of the probabilistic embeddings obtained from our best-performing model trained with *Bing* and *Sentinel-2* imagery, respectively, without any metadata.

The embedding dimension of our probabilistic embeddings is large (512); therefore, in these examples, uncertainty estimates are represented through $||\sigma||_1$ for each sample. We observe that samples with low uncertainty have fewer visible concepts captured in them, suggesting less ambiguity in the types of potential sounds that could be heard at the location. Conversely, for samples with high uncertainty, we usually find denser geographic areas where one would expect to hear multiple types of sounds, leading to higher ambiguity in soundscape mapping.

We also present country-scale uncertainty maps of the USA using PSM's satellite embeddings from both *Bing* and *Sentinel-2*. These maps are shown for zoom levels 1 and 5 in Figure 3. From this figure, we observe that the overall distribution of uncertainty tends to be lower for zoom level 1 compared to zoom level 5. This result is expected because imagery at zoom level 5 covers a larger geographic area, potentially capturing a greater diversity of soundscapes and leading to higher uncertainty in our probabilistic embedding space. Furthermore, a closer examination of the uncertainty values reveals that uncertainty estimates for *Sentinel-2* image embeddings are relatively higher across more locations in the region compared to *Bing* image embeddings. This is expected because for a similar image size, a *Sentinel-2* image with 10m Ground Sampling Distance (GSD) covers a larger area compared to a *Bing* image with 0.6m GSD used in our study.

3 SOUNDSCAPE MAPS

In Figure 4, we present examples of country-scale soundscape maps over the USA. These maps were generated using our best-performing model trained on *Sentinel-2* imagery without any meta-data. In this demonstration, we utilize *Sentinel-2* imagery covering the USA at zoom-level 1. In the figure, for the text query *"Sound of animals on a farm"*, high activation is observed primarily in non-urban areas across the USA. Conversely, for the text query *"Sound of machines in a factory"*, higher activation is concentrated in urban areas near cities, with minimal activation in forested and range-land regions. The use of PSM trained on freely available *Sentinel-2* imagery enables the creation of global-scale soundscape maps.

In Figure 5, we showcase multi-scale soundscape mapping across various geographic regions in the USA. Our objective is to investigate how embeddings and associated similarity scores change with variations in imagery zoom level and imagery source. We generate soundscape maps using *Sentinel-2* satellite image embeddings computed from imagery at zoom levels 1 and 5. To illustrate, we randomly select an audio sample from the *cow* class in the ESC-50 dataset [7] as an example audio query. For the text queries we select *"Sound of children playing in a park"* and *"Sound of machines in a factory"*. For each queries, we analyze the corresponding soundscape maps generated at different zoom levels. In Figure 5, each soundscape map is accompanied by a land cover map [5] of the respective region for reference.

As observed in Figure 5, geographic regions expected to be related to the query demonstrate high similarity scores. For example, for the audio query described by the audio class cow, we can see that urban regions around cities like *Memphis* and *Toledo* have low similarity scores, while rural areas (with greater potential to contain farm animals) exhibit high similarity scores. Similarly, for the text query related to the sound of children playing in a park, as expected, we observe high similarity scores around cities where one would expect to find city parks.

We also observe that for the same query and geographic region, the distribution of similarity scores varies between the two zoom levels. In Figure 5, at zoom level 1, the generated maps appear to be more spatially fine-grained compared to maps generated using satellite imagery at zoom level 5, which appear coarser. Although the number of geolocations and their corresponding satellite imagery is the same for maps at both zoom levels, the coverage area for an image at a higher zoom level is larger. This results in a slower change of high-level visual appearance between the points in the region, leading to closer similarity scores between local points and ultimately producing soundscape maps with lower resolution. This suggests that if we prioritize soundscape maps that retain the semantics of audio concepts at the expense of fine-grained localization capability, we can use satellite imagery at a higher zoom level, which requires fewer images to cover a region of interest. Conversely, for tasks requiring spatially fine-grained soundscape maps, satellite imagery at a lower zoom level may be preferred. This trade-off is fundamental to the multi-scale mapping capability of our framework Probabilistic Soundscape Mapping (PSM).

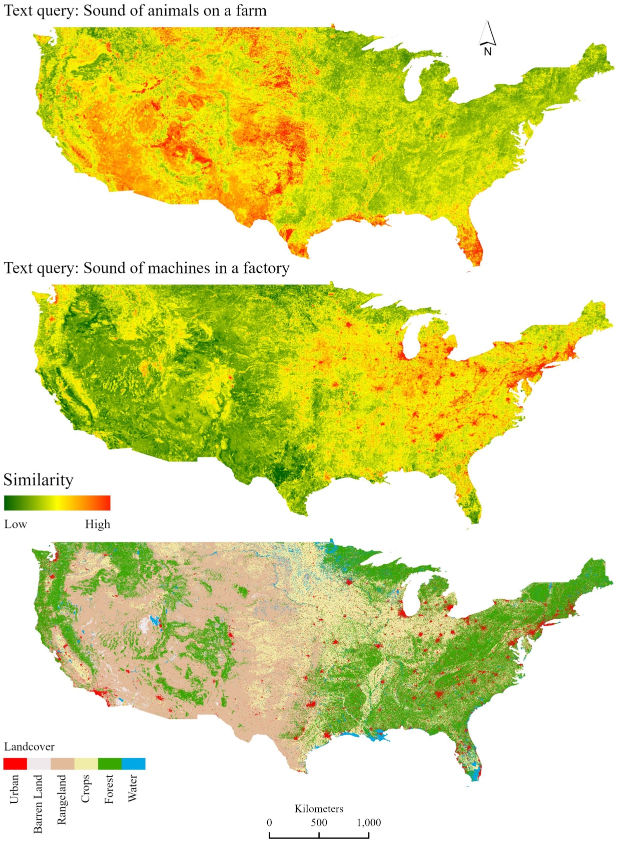

Figure 4: Two soundscape maps of the continental United States, generated from *Sentinel-2* image embeddings, accompanied by a land cover map for reference [5].

REFERENCES

[1] [n. d.]. EOX::Maps, https://tiles.maps.eox.at.
[2] [n. d.]. Freesound, https://freesound.org/.
[3] [n. d.]. iNaturalist, https://www.inaturalist.org.
[4] [n. d.]. Radio aporee: Maps - sounds of the world, https://aporee.org.
[5] EROS. [n. d.]. National Land Cover Database. https://www.usgs.gov/centers/eros/science/national-land-cover-database
[6] Konrad Heidler, Lichao Mou, Di Hu, Pu Jin, Guangyao Li, Chuang Gan, Ji-Rong Wen, and Xiao Xiang Zhu. 2023. Self-supervised audiovisual representation learning for remote sensing data. *International Journal of Applied Earth Observation and Geoinformation* 116 (2023), 103130.
[7] Karol J. Piczak. [n. d.]. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia* (Brisbane, Australia, 2015-10-13). ACM Press, 1015–1018. https://doi.org/10.1145/2733373.2806390
[8] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The New Data in Multimedia Research. *Commun. ACM* 59, 2 (2016), 64–73. http://cacm.acm.org/magazines/2016/2/197425-yfcc100m/fulltext

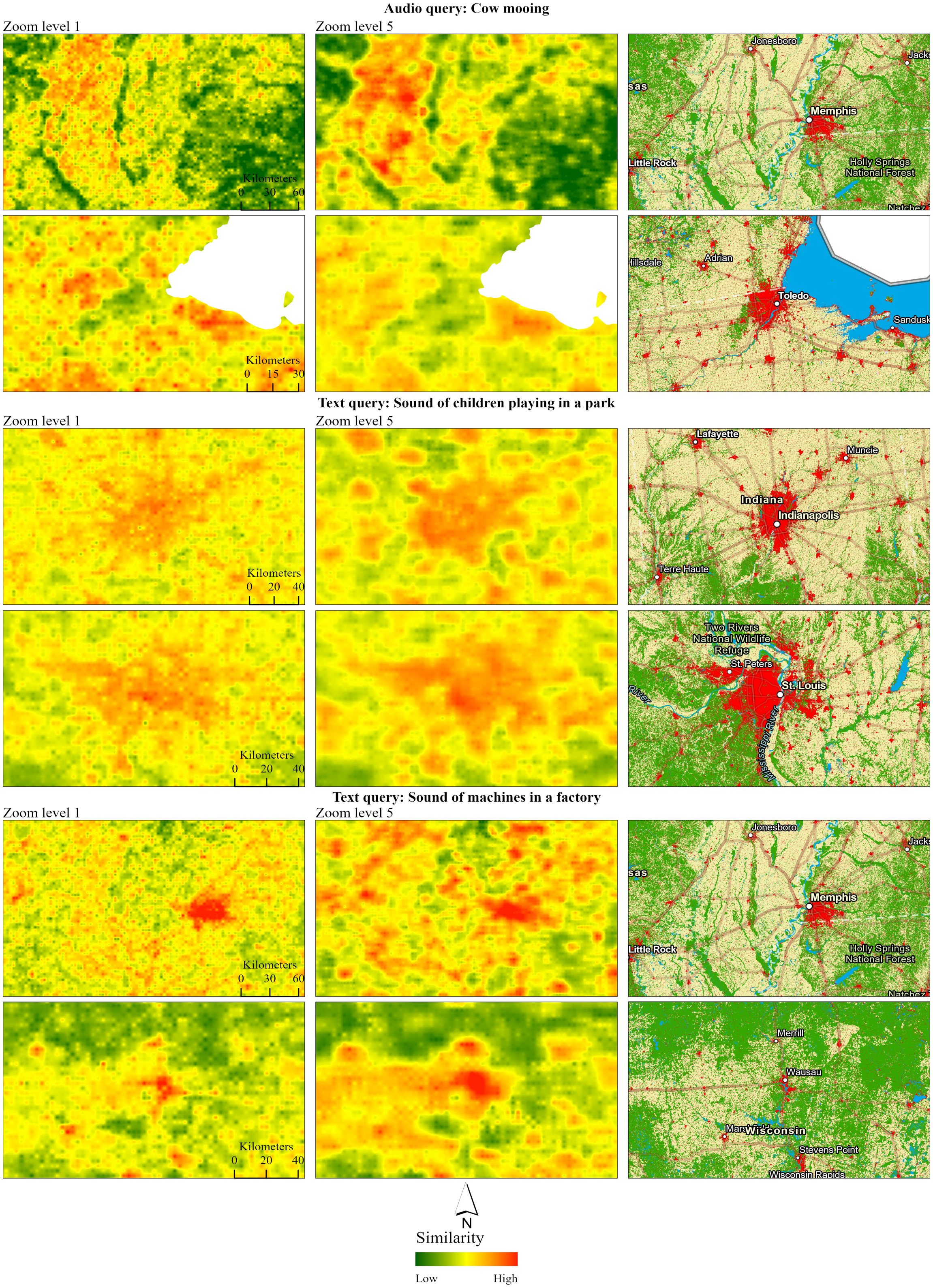

Figure 5: Soundscape maps over smaller geographic areas, computed using similarity scores between respective queries and embeddings from *Sentinel-2* satellite imagery at two zoom levels.