

PSM: Learning Probabilistic Embeddings for Multi-scale Zero-Shot Soundscape Mapping

Subash Khanal

Washington University in St. Louis
Saint Louis, United States
k.subash@wustl.edu

Aayush Dhakal

Washington University in St. Louis
Saint Louis, United States
a.dhakal@wustl.edu

Eric Xing

Washington University in St. Louis
Saint Louis, United States
e.xing@wustl.edu

Zhexiao Xiong

Washington University in St. Louis
Saint Louis, United States
x.zhexiao@wustl.edu

Nathan Jacobs

Washington University in St. Louis
Saint Louis, United States
jacobsn@wustl.edu

Srikumar Sastry

Washington University in St. Louis
Saint Louis, United States
s.sastry@wustl.edu

Adeel Ahmad

Washington University in St. Louis
Saint Louis, United States
aadeel@wustl.edu

ABSTRACT

A soundscape is defined by the acoustic environment a person perceives at a location. In this work, we propose a framework for mapping soundscapes across the Earth. Since soundscapes involve sound distributions that span varying spatial scales, we represent locations with multi-scale satellite imagery and learn a joint representation among this imagery, audio, and text. To capture the inherent uncertainty in the soundscape of a location, we design the representation space to be probabilistic. We also fuse ubiquitous metadata (including geolocation, time, and data source) to enable learning of spatially and temporally dynamic representations of soundscapes. We demonstrate the utility of our framework by creating large-scale soundscape maps integrating both audio and text with temporal control. To facilitate future research on this task, we also introduce a large-scale dataset, GeoSound, containing over 300k geotagged audio samples paired with both low- and high-resolution satellite imagery. We demonstrate that our method outperforms the existing state-of-the-art on both GeoSound and the existing SoundingEarth dataset. Our dataset and code is available at <https://github.com/mvrl/PSM>.

KEYWORDS

Soundscape Mapping; Audio Visual Learning; Probabilistic Representation Learning

1 INTRODUCTION

Soundscape mapping involves understanding the relationship between locations on Earth and the distribution of sounds at those

locations. The soundscape of an area strongly correlates with psychological and physiological health [27]. Therefore, soundscape maps are valuable tools for environmental noise management and urban planning [19, 30, 32]. Additionally, technologies such as augmented reality and navigation systems can use soundscape mapping to provide immersive experiences.

Traditionally, soundscape mapping involves predicting a set of acoustic indicators (e.g., sound pressure, loudness) to descriptors (e.g., pleasant, eventful) [16, 18, 28]. This approach limits understanding of the acoustic scene and relies on crowd-sourced data [6, 35], often available only for densely populated areas. As a result, traditional methods generate sparse soundscape maps that lack generalizability and are unsuitable for creating global maps.

To address the limitations of traditional soundscape mapping, we adopt a formulation where, given a specific location, the task is to train a machine learning model that directly predicts the sound distribution likely to be encountered at that location. We represent each location with a satellite image centered around it. This approach enables the generalization of soundscape mapping beyond locations explicitly included in the training data.

We approach the soundscape mapping problem from the perspective of multimodal representation learning to design a shared embedding space between audio and satellite imagery at the recorded location of the audio. This learning strategy aims to bring positive audio-satellite image pairs closer while pushing negative pairs farther apart in the embedding space. Ultimately, the multimodal embedding space can be employed to generate soundscape maps by computing similarity scores between the query and the satellite image set covering the geographic region of interest.

Soundscape mapping is inherently uncertain, as multiple types of sounds can come from one location, and a specific sound can originate from multiple locations. Paired location and audio data often contain pseudo-positives, sample pairs labeled as negatives but semantically similar to positives. Methods that learn deterministic representations of sound and satellite imagery ignore this uncertainty. To address this, we propose a probabilistic multi-modal

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0686-8/24/10

<https://doi.org/10.1145/3664647.3681620>

embedding space for audio, satellite imagery, and textual descriptions [12]. To account for potential false negatives, we identify pseudo-positive matches during training [12].

Satellite images of audio capture locations can be obtained at different spatial resolutions, with ground area coverage increasing as zoom levels increase. To create large-scale soundscape maps, we aim to learn an embedding space that models these differences in spatial resolution of satellite imagery. Therefore, we modify zero-shot soundscape mapping to multi-scale zero-shot soundscape mapping, allowing ground-level sounds to be mapped with satellite imagery at different zoom levels. We achieve this by learning a shared satellite image encoder across zoom levels using Ground-Sample Distance Positional Embedding (GSDPE) [38].

Our modalities of interest, satellite imagery, audio, and text, often have associated metadata that convey meaningful information (such as latitude and longitude or the source of an audio sample). We propose to fuse such metadata: location, time, and source from which the audio was collected, into our framework. We demonstrate that such information increases the discriminative power of our embedding space and allows the creation of soundscape maps conditioned on dynamic metadata settings during inference.

The most closely related prior work [26] in soundscape mapping was trained on limited data (~ 35k samples) from the *SoundingEarth* dataset [21]. To advance research in this area, we curated a new large-scale dataset, *GeoSound*, collecting geotagged audios from four different sources, we increased the dataset size by six-fold to over 300k samples. We use *GeoSound* to train our framework that advances the state-of-the-art in zero-shot soundscape mapping by learning a probabilistic, scale-aware, and metadata-aware joint multimodal embedding space. Moreover, we demonstrate the capability of the proposed framework in the creation of temporally dynamic soundscape maps.

The main contributions of our work are as follows:

- We introduce a new large-scale dataset containing over 300k geotagged audios paired with high-resolution (0.6m) and low-resolution (10m) satellite imagery.
- We learn a metadata-aware, probabilistic embedding space between satellite imagery, audio, and textual audio description for zero-shot multi-scale soundscape mapping.
- We demonstrate the utility of our framework (PSM: Probabilistic Soundscape Mapping) in creating large-scale soundscape maps created by querying our learned embedding space with audio or text.

2 RELATED WORKS

2.1 Audio Visual Learning

An intricate relationship exists between the audio and visual attributes of a scene. Utilizing this relationship, there have been numerous works in the field of audio-visual learning. [10, 21, 22, 24, 26, 34, 39, 47, 48]. Owens *et al.* [34] have demonstrated that encouraging the models to predict sound characteristics of a scene allows them to learn richer representations useful for visual recognition tasks. Hu *et al.* [22] proposed to learn from audio and images to solve the task of aerial scene recognition. Relatively closer to the formulation of our work, Salem *et al.* [39] proposed to learn a shared feature space between satellite imagery, ground-level concepts, and

audio, which allowed them to predict sound cluster distribution across large geographic regions. Recently, Khanal *et al.* [26] proposed the learning of a tri-modal embedding space to map satellite imagery with the most likely audio at a location.

2.2 Deterministic Contrastive Learning

The contrastive learning paradigm [29, 37, 41, 44] has significantly advanced state-of-the-art multimodal learning capabilities through rich cross-modal supervision. In the pursuit of advancing contrastive learning approaches for audio and text, Elizalde *et al.* [15] and Wu *et al.* [46] have developed a Contrastive Language-Audio Pretraining (CLAP) framework, showcasing strong zero-shot capabilities. Wav2CLIP [45] distills information learned from CLIP to create a joint image-audio embedding space. AudioCLIP [20] extends contrastive learning to audio, image, and text, exhibiting impressive performance across various downstream tasks. Recently, Heidler *et al.* proposed learning a shared representation space between audio and corresponding satellite imagery for use in various downstream tasks in remote sensing. Similarly, Khanal *et al.* [26] utilized the *SoundingEarth* dataset [21] to train a multimodal embedding space using a deterministic contrastive loss [33] and used this model for zero-shot soundscape mapping.

2.3 Probabilistic Contrastive Learning

In our formulation of soundscape mapping, the satellite image provided as location context captures a geographic area containing many sound sources. As such, deterministic contrastive learning approaches cannot capture the inherent ambiguity in the mapping from satellite image to sound, as any sample can only be represented by a single point in the embedding space. This limitation can be addressed by representing embeddings probabilistically [8, 9, 12, 13, 23, 25, 31, 40, 43]. In other words, each sample in probabilistic embedding space is represented by a probability distribution whose parameters are learned, usually by a neural network.

Probabilistic Cross-Modal Embeddings (PCME) [13] represents samples as gaussian distributions in the embedding space and trains their framework using a contrastive loss between the sample distributions computed by Monte-Carlo sampling. Chun [12] later introduced PCME++, which improves PCME by using a closed-form distance formulation, eliminating the need for sampling. We adopt PCME++ to learn a probabilistic embedding space for audio, its textual description, and multi-scale satellite imagery.

3 METHODS

3.1 Dataset Creation

Prior work in zero-shot soundscape mapping [26] has utilized the *SoundingEarth* dataset [21], which contains approximately 50k geotagged audios paired with corresponding satellite imagery. To facilitate research on training large-scale models with a rich representation space for soundscape mapping, we have expanded the size of the dataset 6-fold by creating a dataset containing 309 019 geo-tagged audios from four different sources: *iNaturalist* [4], *yfcc-video* [42], *aporee* [5], and *freesound* [2], each contributing 114 603, 96 452, 49 284, and 48 680 samples respectively. We pair these geotagged audios with their corresponding *Sentinel-2-cloudless* imagery [1] with 10m GSD and *Bing* imagery with 0.6m GSD.

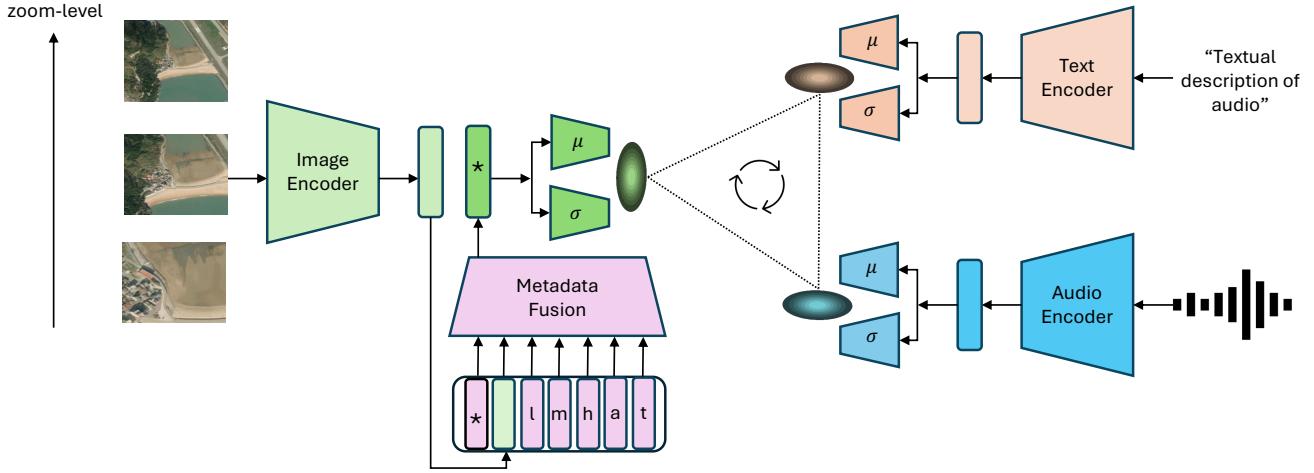


Figure 1: Our proposed framework, Probabilistic Soundscape Mapping (PSM), combines image, audio, and text encoders to learn a probabilistic joint representation space. Metadata, including geolocation (l), month (m), hour (h), audio-source (a), and caption-source (t), is encoded separately and fused with image embeddings using a transformer-based metadata fusion module. For each encoder, μ and σ heads yield probabilistic embeddings, which are used to compute probabilistic contrastive loss.

In the prior work, GeoCLAP [26], samples were randomly split between train/validation/test sets for training and evaluating their models. We observed that such a data split strategy leads to the issue of data leakage where evaluation data samples come from the same set of locations present in the training set, preventing the evaluation of the generalizability of a model to truly unseen locations. To address this, we divide the world into $1^\circ \times 1^\circ$ non-overlapping cells where each cell containing some samples is randomly assigned to either train/validation/test set. Our dataset contains 294 019/5000/10 000 samples in the train/validation/test sets. We also employ our split strategy on the *SoundingEarth* dataset with a cell size of $10\text{km} \times 10\text{km}$. This strategy resulted in 41 469/3242/5801 samples in train/validation/test sets. Details of our dataset and split strategy are in the supplemental material.

3.2 Approach

This section describes our framework (PSM) for learning a metadata-aware, probabilistic, and tri-modal embedding space for multi-scale zero-shot soundscape mapping.

Figure 1 presents an overview of the PSM framework, which comprises an image encoder, metadata fusion module, text encoder, and audio encoder. The scale-aware image encoder converts multi-scale satellite imagery into a d -dimensional representation. The transformer-based metadata fusion module integrates metadata (including location, month, time, audio source, and text source) with the image representation, generating a metadata-aware probabilistic image representation. Other modality-specific encoders produce probabilistic embeddings for text and audio. PSM aims to map tuples of satellite imagery, audio, and text into a shared probabilistic representation space.

Given a geotagged audio X_k^a , textual description of the audio X_k^t , and a satellite image at a given location viewed at a zoom level l (an integer between 1 and some maximum zoom level L) $X_{k,l}^i$,

$(X_k^a, X_k^t, X_{k,l}^i)$ is the k -th audio-text-image triplet. PSM is trained over the aggregation of all available triplets.

We use modality-specific transformer-based encoders followed by their respective linear projection modules to obtain representations $(h_k^a, h_k^t, h_{k,l}^i)$ with same dimension d ,

$$h_k^a = g_{\text{audio}}(f_{\text{audio}}(X_k^a)) \quad (1)$$

$$h_k^t = g_{\text{text}}(f_{\text{text}}(X_k^t)) \quad (2)$$

$$h_{k,l}^i = g_{\text{image}}(f_{\text{image}}(X_{k,l}^i, l_k)) \quad (3)$$

where $(f_{\text{audio}}, g_{\text{audio}})$, $(f_{\text{text}}, g_{\text{text}})$, $(f_{\text{image}}, g_{\text{image}})$ are (encoder, projection-module) pairs producing d dimensional embeddings: h_k^a , h_k^t , and $h_{k,l}^i$ for audio, text, and satellite image with zoom-level l_k respectively.

We use GSDPE [38] to encode the position and scale of each patch of satellite imagery at zoom-level l to learn scale-aware representations of multi-scale satellite imagery,

$$v_{l,x}(pos, 2i) = \sin\left(\frac{g * l}{G}\right) \frac{pos}{10000 \frac{2i}{d}} \quad (4)$$

$$v_{l,y}(pos, 2i + 1) = \cos\left(\frac{g * l}{G}\right) \frac{pos}{10000 \frac{2i}{d}} \quad (5)$$

where pos is the position of the image patch along the given axis (x or y), i is the feature dimension index, l is the zoom-level of the image, g is the GSD of the image, and G is the reference GSD.

As discussed before, we are interested in learning metadata-aware representation space. Therefore, we fuse four different components of metadata (geolocation, month, hour, audio-source, caption-source) with the satellite image embedding (h_k^i) and obtain a metadata-conditioned image embedding ($h_k^{i'}$),

$$h_k^{i'} = g_{\text{meta}}(h_k^i, \text{metadata}) \quad (6)$$

where g_{meta} is a transformer based metadata fusion module of our framework, h_k' is the embedding corresponding to the learnable special token (*) fed into g_{meta} .

To learn a probabilistic embedding space, we define the embedding of a given modality (r) as a normally distributed random variable, $Z_r \sim N(\mu_r, \sigma_r)$. We employ a closed-form probabilistic contrastive loss [12] between all three pairs of embeddings. For any two modalities p and q , the closed-form sampled distance (CSD) as defined in PCME++ [12] is as follows:

$$d(Z_p, Z_q) = \|\mu_p - \mu_q\|_2^2 + \|\sigma_p^2 + \sigma_q^2\|_1. \quad (7)$$

In our implementation, we pass our modality-specific representations, h_k^a , h_k^t , and h_k' , through heads for μ and $\log(\sigma^2)$ of the gaussian distribution representing our samples.

This distance is used to compute matching loss between the pairs as shown in Equation 8. For positive (certain) matches, to reduce distance, μ embeddings are pulled closer together while minimizing σ embeddings in $\|\sigma_p^2 + \sigma_q^2\|_1$. Similarly, for negative (uncertain) match, to increase the distance, μ embeddings are pushed farther while maximizing σ embeddings in $\|\sigma_p^2 + \sigma_q^2\|_1$.

Based on the distance function defined in Equation 7, we can define the probabilistic matching objective function as follows:

$$\mathcal{L}_m = -[w_{pq} \log(\text{sigmoid}(-a.d(Z_p, Z_q) + b)) + (1 - w_{pq}) \log(\text{sigmoid}(a.d(Z_p, Z_q) - b))] \quad (8)$$

where $w_{pq} \in \{0, 1\}$ is the matching indicator between p and q . a and b are learnable scalar parameters. \mathcal{L}_m (\mathcal{L}_{match}) is computed for all sample pairs in the mini-batch.

Soundscape mapping is inherently a one-to-many matching problem. Given a satellite image at a location, there may be multiple sounds that are likely to be heard there. Therefore, if we were to simply assign w_{pq} as 0 or 1 for our dataset's negative and positive matches, we would lose the opportunity to learn from the potentially numerous false negatives. Therefore, we adopt a similar strategy of learning from pseudo-positives, as formulated by Chun [12]. In this approach, for a positive match (p, q) , we consider q' as a pseudo-positive match ($pseudo-m$) with q if $d(Z_p, Z_{q'}) \leq d(Z_p, Z_q)$. Finally, the objective function for a pair of modalities (p, q) is as follows:

$$\mathcal{L}_{p,q} = \mathcal{L}_m + \alpha \mathcal{L}_{pseudo-m} + \beta \mathcal{L}_{VIB} \quad (9)$$

where α and β control for the contribution of pseudo-match loss and Variational Information Bottleneck (VIB) loss [7], respectively. We use \mathcal{L}_{VIB} as a regularizer to reduce overfitting, preventing the collapse of σ to 0.

To learn a tri-modal embedding space for zero-shot soundscape mapping, using Equation 9, we separately compute loss for all three pairs of modalities: audio-text (a, t), audio-image (a, i), and image-text(i, t). Finally, the overall objective function to train PSM is as follows:

$$\mathcal{L} = \mathcal{L}_{a,t} + \mathcal{L}_{a,i} + \mathcal{L}_{i,t}. \quad (10)$$

4 EXPERIMENTAL DETAILS

Audio/Text Processing: We use pre-trained models for the audio and text modalities and their respective input processing pipelines

hosted on HuggingFace. Specifically, for audio, we extract the audio spectrogram using the ClapProcessor wrapper for the pre-trained CLAP [46] model clap-htsat-fused with default parameters: `feature_size=64`, `sampling_rate=48000`, `hop_length=480`, `fft_window_size=1024`. CLAP uses a feature fusion strategy [46] to pre-process variable length sounds by extracting a spectrogram of randomly selected 3 d -second audio slices and the spectrogram of the whole audio down-sampled to 10s. We choose $d = 10$ s in our experiments. Apart from the text present in the metadata, we also obtain a textual description of audio from a recent SOTA audio captioning model, Qwen-sound [11], and use the captioning model's output only if it passes CLAP-score [46] based quality check. For the textual descriptions of audio in our data, we adopt the similar text processing as performed by CLAP [46] and tokenize our text using RobertaTokenizer with `max_length=128`.

Satellite image processing: Our framework is trained with satellite images at different zoom levels $l \in \{1, 3, 5\}$. To obtain this data, we first downloaded a large tile of images with size $(Lh) \times (Lw)$. We obtained high-resolution 0.6m GSD imagery with a tile size of 1500×1500 from *Bing* and low-resolution 10m GSD imagery with a tile size of 1280×1280 from *Sentinel-2-cloudless*. To get an image at zoom-level l , we center crop from the original tile with a crop size of $(lh) \times (lw)$ and then resize it to an $h \times w$ image, where (h, w) is $(256, 256)$ for *Sentinel-2* imagery and $(300, 300)$ for *Bing* imagery. This way, we can simulate the effect of change in coverage area as the zoom-level changes while effectively keeping constant input image size for training. During training, we randomly sample l from a set $\{1, 3, 5\}$ for each image instance. Then, for the zoom-transformed image, we perform *RandomResizedCrop* with parameters: `{input_size=224, scale=(0.2, 1.0)}` followed by a *RandomHorizontalFlip* while only extracting a 224×224 center crop of the image at the desired zoom-level l for evaluation.

Metadata Fusion: To fuse metadata into our framework, we first separately project the metadata components into 512-dimensional space using linear layers and concatenate them with the satellite image embedding from the image encoder and a learnable special token. Finally, the set of tokens is fed into a lightweight transformer-based module containing only 3 layers. The output of this module is further passed through heads for μ and $\log(\sigma^2)$ of the Gaussian distribution representing metadata-conditioned image embeddings. To avoid overfitting to the metadata, we independently drop each metadata component at the rate of 0.5 during training.

Training: We initialize encoders from released weights of pre-trained models, SatMAE [14] for satellite imagery and CLAP [46] for audio and text. We chose d , the dimensionality of our embeddings, to be 512. For regularization, we set the weight decay to 0.2. Our training `batch_size` was 128. We use Adam as our optimizer, with the initial learning rate set to $5e-5$. To schedule the learning rate, we use cosine annealing with warm-up iterations of 5k for experiments with *GeoSound* and 2k for experiments with *SoundingEarth*.

Baseline: We use GeoCLAP [26], a SOTA zero-shot soundscape mapping model, as a baseline for evaluation. GeoCLAP is contrastively trained using the *infoNCE* [33] loss between three modality pairs: image-audio, audio-text, and image-text.

Metrics: We evaluate on two datasets: *GeoSound*, and *SoundingEarth*. We use Recall@10% and the Median Rank of the ground truth as our

method	loss	meta/train	text/eval	meta/eval	ZL	I2A R@10%	I2A MdR	A2I R@10%	A2I MdR
GeoCLAP	infoNCE	✗	✗	✗	1	0.399	1500	0.403	1464
GeoCLAP	infoNCE	✗	✓	✗	1	0.577	712	0.468	1141
ours	infoNCE	✓	✓	✓	1	0.709	462	0.871	241
ours	PCME++	✗	✗	✗	1	0.423	1401	0.428	1344
ours	PCME++	✓	✗	✓	1	0.828	261	0.829	248
ours	PCME++	✓	✓	✓	1	0.901	113	0.943	100
GeoCLAP	infoNCE	✗	✗	✗	3	0.408	1441	0.420	1389
GeoCLAP	infoNCE	✗	✓	✗	3	0.577	707	0.483	1056
ours	infoNCE	✓	✓	✓	3	0.708	462	0.875	235
ours	PCME++	✗	✗	✗	3	0.440	1302	0.443	1266
ours	PCME++	✓	✗	✓	3	0.827	266	0.832	250
ours	PCME++	✓	✓	✓	3	0.900	114	0.945	102
GeoCLAP	infoNCE	✗	✗	✗	5	0.409	1428	0.421	1373
GeoCLAP	infoNCE	✗	✓	✗	5	0.581	698	0.489	1036
ours	infoNCE	✓	✓	✓	5	0.709	461	0.875	238
ours	PCME++	✗	✗	✗	5	0.440	1302	0.448	1279
ours	PCME++	✓	✗	✓	5	0.821	281	0.826	261
ours	PCME++	✓	✓	✓	5	0.896	115	0.941	107

Table 1: Experimental results for models trained on the GeoSound dataset with satellite imagery from *Bing*.

method	loss	meta/train	text/eval	meta/eval	ZL	I2A R@10%	I2A MdR	A2I R@10%	A2I MdR
GeoCLAP	infoNCE	✗	✗	✗	1	0.459	1179	0.465	1141
GeoCLAP	infoNCE	✗	✓	✗	1	0.546	827	0.553	804
ours	infoNCE	✓	✓	✓	1	0.722	497	0.860	247
ours	PCME++	✗	✗	✗	1	0.474	1101	0.485	1061
ours	PCME++	✓	✗	✓	1	0.802	294	0.804	283
ours	PCME++	✓	✓	✓	1	0.872	142	0.940	104
GeoCLAP	infoNCE	✗	✗	✗	3	0.454	1200	0.456	1197
GeoCLAP	infoNCE	✗	✓	✗	3	0.542	840	0.555	790
ours	infoNCE	✓	✓	✓	3	0.722	491	0.856	248
ours	PCME++	✗	✗	✗	3	0.479	1086	0.487	1042
ours	PCME++	✓	✗	✓	3	0.795	306	0.800	290
ours	PCME++	✓	✓	✓	3	0.870	150	0.940	104
GeoCLAP	infoNCE	✗	✗	✗	5	0.458	1194	0.457	1184
GeoCLAP	infoNCE	✗	✓	✗	5	0.542	835	0.554	791
ours	infoNCE	✓	✓	✓	5	0.719	497	0.852	252
ours	PCME++	✗	✗	✗	5	0.459	1172	0.465	1138
ours	PCME++	✓	✗	✓	5	0.794	316	0.794	299
ours	PCME++	✓	✓	✓	5	0.868	156	0.935	109

Table 2: Experimental results for models trained on the GeoSound dataset with satellite imagery from *Sentinel-2*.

evaluation metrics. Recall@10% (R@10%) is defined as the proportion of queries for which the ground-truth match is found within the top 10% of the returned ranked retrieval list. Median Rank (MdR) is defined as the median of the positions at which the ground-truth matches appear in the ranked retrieval list. We denote image-to-audio as I2A and audio-to-image as A2I throughout the paper. To

assess the effectiveness of text embeddings in cross-modal retrieval between satellite images and audio, we also evaluate an experimental setting where, during inference, we add the corresponding text embedding to the query embedding during retrieval.

method	loss	meta/train	text/eval	meta/eval	I2A R@10%	I2A MdR	A2I R@10%	A2I MdR
GeoCLAP	infoNCE	✗	✗	✗	0.454	667	0.449	694
GeoCLAP	infoNCE	✗	✓	✗	0.523	533	0.470	641
ours	infoNCE	✓	✓	✓	0.519	548	0.491	596
ours	PCME++	✗	✗	✗	0.514	547	0.518	543
ours	PCME++	✓	✗	✓	0.563	454	0.569	447
ours	PCME++	✓	✓	✓	0.690	264	0.608	371

Table 3: Experimental results for models trained on the SoundingEarth dataset.

imagery	latlong	month	time	audio source	text source	I2A R@10%	I2A MdR	A2I R@10%	A2I MdR
Sentinel-2	✓	✗	✗	✗	✗	0.512	946	0.516	923
Sentinel-2	✗	✓	✗	✗	✗	0.501	988	0.511	941
Sentinel-2	✗	✗	✓	✗	✗	0.548	825	0.574	717
Sentinel-2	✗	✗	✗	✓	✗	0.749	407	0.757	389
Sentinel-2	✗	✗	✗	✗	✓	0.483	1080	0.492	1022
Bing	✓	✗	✗	✗	✗	0.539	822	0.557	764
Bing	✗	✓	✗	✗	✗	0.464	1153	0.485	1068
Bing	✗	✗	✓	✗	✗	0.516	937	0.547	823
Bing	✗	✗	✗	✓	✗	0.722	469	0.733	447
Bing	✗	✗	✗	✗	✓	0.448	1250	0.466	1140

Table 4: Metadata ablation to evaluate the impact of individual metadata components on the best model’s performance. The best model for each imagery type was trained with all metadata types and can handle multiple zoom levels. The results presented are specific to satellite imagery at zoom level 1, with no text embeddings added to the query embeddings during inference.

5 RESULTS

5.1 Cross-Modal Retrieval with Bing

Table 1 presents our retrieval evaluation of PSM trained on the *GeoSound* dataset using *Bing* satellite imagery. Our approach outperforms the state-of-the-art baseline [26] for cross-modal retrieval between satellite imagery and audio, and vice versa. SatMAE [14] with GSDPE is utilized to encode the zoom level of the satellite imagery for both the baseline and our models. This enables our satellite image encoder to remain invariant to zoom-level changes, achieving consistent performance across all zoom levels. We observe that learning a probabilistic embedding space using PCME++ loss alone enhances the baseline performance from 0.399 to 0.423, 0.408 to 0.440, and 0.409 to 0.440 for zoom levels 1, 3, and 5, respectively. In addition to the objective function, we also experimented with the inclusion of metadata during training and inference. As anticipated, the model’s performance, when trained and evaluated with both text and metadata, is notably improved, enhancing image-to-audio R@10% from the baseline score of 0.577 to 0.901, 0.577 to 0.900, and 0.581 to 0.896 for zoom levels 1, 3, and 5, respectively. A similar trend is observed for audio-to-image retrieval.

5.2 Cross-Modal Retrieval with Sentinel-2

Table 2 presents the evaluation results of PSM trained on the *GeoSound* dataset using *Sentinel-2* satellite imagery. Similar to results with *Bing* imagery, we observe consistent performance across various zoom levels, indicating the robustness of our framework in

extracting valuable information irrespective of the coverage area of input satellite imagery. By employing PCME++ loss in training our framework, we note an enhancement in the baseline performance from 0.459 to 0.474 for zoom level 1. Overall, PSM trained with *Sentinel-2* imagery and metadata, and evaluated using both metadata and text during inference, significantly improved the baseline score from 0.546 to 0.872, 0.542 to 0.870, and 0.542 to 0.868 for zoom levels 1, 3, and 5, respectively. A similar trend is observed for audio-to-image retrieval. Moreover, the high performance of PSM on *Sentinel-2* imagery at zoom level 5 enables the efficient creation of large-scale soundscape maps using freely available *Sentinel-2* imagery while requiring fewer images to download.

5.3 Cross-Modal Retrieval on SoundingEarth

Table 3 presents the evaluation results of PSM trained on the *SoundingEarth* dataset [21] with its original 0.2m GSD *GoogleEarth* imagery. For the *SoundingEarth* dataset, our models are exclusively trained and evaluated on zoom level 1. Similar to the performance observed on the *GeoSound* dataset, we witness gain in performance with our approach of learning a metadata-aware probabilistic embedding space. Specifically, by training with the PCME++ objective instead of the *infoNCE* loss, we note an improvement in the score from 0.454 to 0.514. This performance further elevates to 0.563 when metadata is incorporated and reaches 0.690 when both metadata and text are utilized during inference. We observe similar trends for audio-to-image retrieval as well.

Text query: Sound of insects

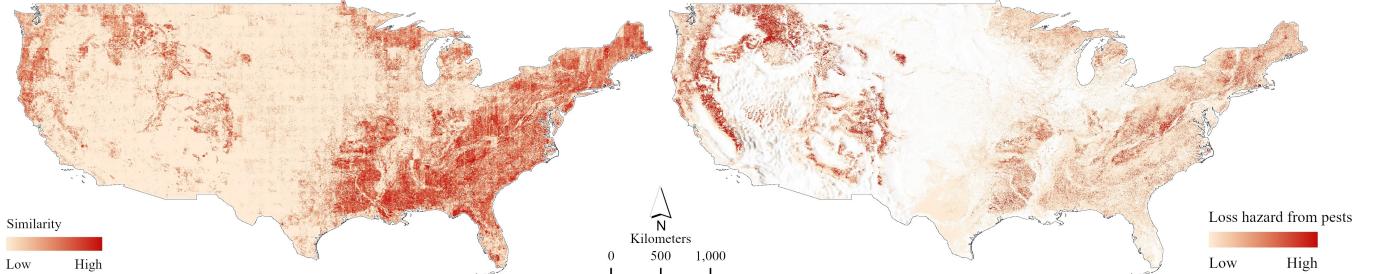
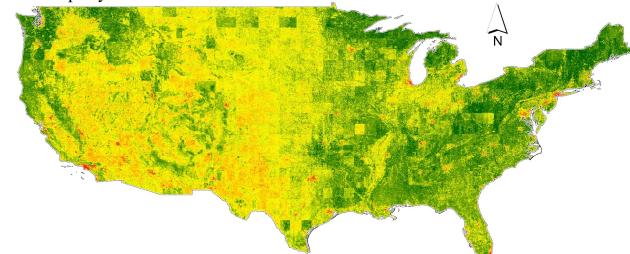


Figure 2: Soundscape Map of the USA for a textual query *Sound of insects*, compared with a reference map [3] indicating the risk of pest-related hazard.

Audio query: Car horn



Text query: Sound of chirping birds

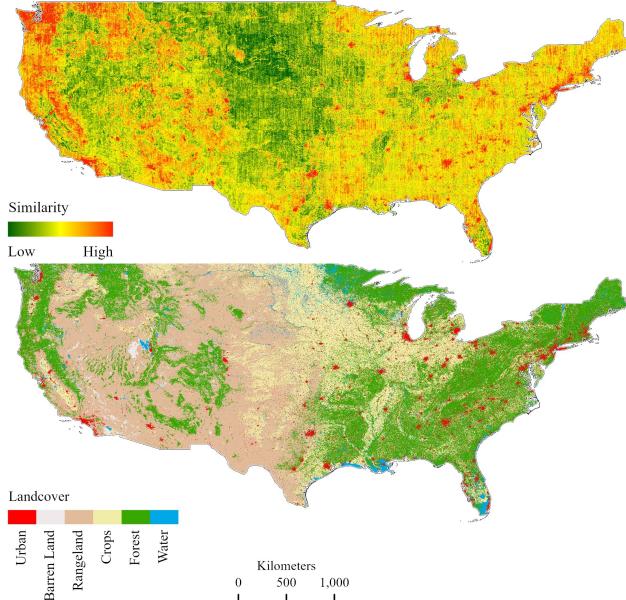


Figure 3: Two soundscape maps of the continental United States, generated using different query types, with a land cover map [17] for reference.

5.4 Effect of Metadata

Our experimental results reveal a significant enhancement in the model’s performance when metadata is integrated into both training and inference. For comparison, as illustrated in Table 1, PSM trained with *Bing* imagery without any metadata achieved an I2A

R@10% of 0.423, whereas with all metadata included, it reached 0.828. A similar trend is seen for experiments with *Sentinel-2* imagery. PSM is designed such that individual metadata components are independently masked out with a rate of 0.5. Therefore, during inference, we can evaluate PSM by dropping any combination of metadata components.

In Table 4, we present the ablation of different metadata components to evaluate the impact of individual metadata components in PSM’s learning framework. We perform this ablation on our best-performing models trained on the *GeoSound* dataset with *Sentinel-2* and *Bing* imagery. The results in Table 4 exclude the use of text during cross-modal retrieval and utilize satellite imagery at zoom level 1 during inference. These results highlight two major findings. First, all of the metadata components contribute to the overall improvement of PSM’s performance. Second, among all of the metadata components, audio-source had the most significant impact. This suggests that biases in different audio hosting platforms are encoded into the learning framework, improving cross-modal retrieval and enabling soundscape maps conditioned on the expected audio type from specific platforms.

5.5 Generating Country-Level Soundscape Maps

We demonstrate PSM’s capability to generate large-scale soundscape maps using audio and text queries. We acquired 0.6 m GSD 1500 × 1500 image tiles encompassing the entire USA from *Bing*.

Using our top-performing model’s image encoder, we pre-compute embeddings for each image at zoom level 1. During inference, these are combined with desired metadata embeddings via the model’s metadata fusion module to obtain metadata-conditioned probabilistic embeddings for the entire region.

We use modality-specific encoders to obtain probabilistic embeddings for audio or text queries. To compute similarity scores between image embeddings and query embeddings, we utilize Equation 7 as detailed in our paper. These scores are then used to produce large-scale soundscape maps, as shown in Figures 2 and 3.

6 DISCUSSION

Figure 2 depicts a soundscape map generated for the textual query “*Sound of insects*”, accompanied by the following metadata: {audio source: iNaturalist, month: May, time: 8 pm}. Notably,

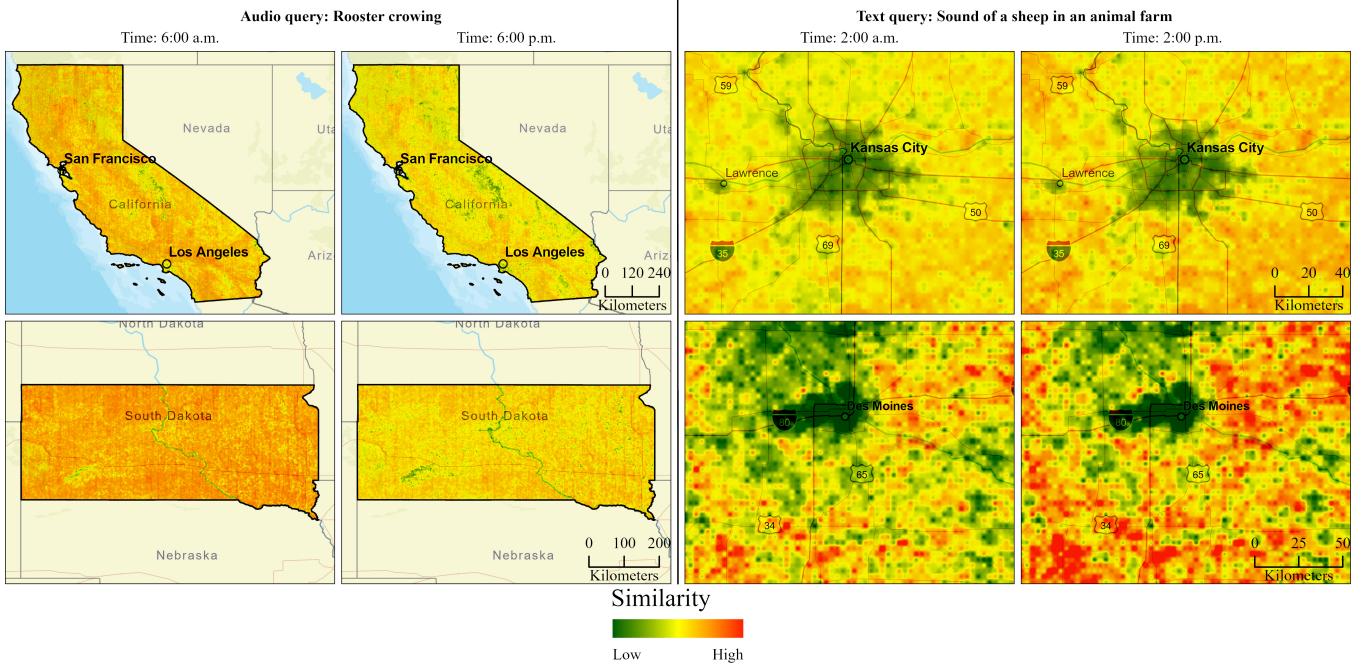


Figure 4: Temporally dynamic soundscape maps created by querying our model for different geographic areas.

this soundscape map exhibits a strong correlation with an available reference map [3], which shows potential pest hazards across the continental United States. Figure 3 showcases two soundscape maps: one for an audio query of *car horn* with the metadata {audio source: yfcc, month: May, time: 10 am}, and another for a textual query “*Sound of chirping birds*.” with metadata: {audio source: iNaturalist, month: May, time: 10 am}. Both maps can be compared with a land cover map [17]. As expected, for the car horn query, higher activation is observed in most major US cities, while for chirping birds, increased activation is observed around both urban areas and forests.

We also note that the soundscape of any geographic region evolves predictably over a day. Therefore, the hour of the day is one of the important metadata components fused into our framework. In addition to contributing to increased performance, temporal understanding fused into our embedding space allows us to create temporally dynamic soundscape maps across any geographic region, as demonstrated in Figure 4. The similarity scores used for these soundscape maps were normalized consistently for a region across time. We display state-level temporally dynamic soundscape maps for an audio query: *Rooster crowing* with metadata: {audio source: aporee, month: May, time: 6 am} vs. {audio source: aporee, month: May, time: 6 pm}. We observe that for both states, higher activation for the *rooster crowing* audio query is seen on the soundscape map at 6 am. We also showcase city-level temporally dynamic soundscape maps for a text query “*Sound of a sheep in an animal farm*.” We can observe that for areas around both cities, Kansas City and Des Moines, very low activation is present. Additionally, higher activation is observed at 2 pm than at 2 am, which is expected. These demonstrations highlight the ability of our

model to create semantically meaningful and temporally consistent soundscape maps across any geographic regions of interest.

7 CONCLUSION

Our work introduces a framework for learning probabilistic trimodal embeddings for the task of multi-scale zero-shot soundscape mapping. To advance research in this direction, we have developed a new large-scale dataset that pairs geotagged audio with high and low-resolution satellite imagery. By utilizing a probabilistic trimodal embedding space, our method surpasses the state-of-the-art while also providing uncertainty estimates for each sample. Furthermore, we have designed our framework to be metadata-aware, resulting in a significant improvement in cross-modal retrieval performance. Additionally, it enables the creation of dynamic soundscape maps conditioned on different types of metadata. The combination of enhanced mapping performance, uncertainty estimation, and a comprehensive understanding of spatial and temporal dynamics positions our framework as an effective solution for zero-shot multi-scale soundscape mapping.

REFERENCES

- [1] [n. d.]. EO::Maps, <https://tiles.maps.eox.at>.
- [2] [n. d.]. Freesound, <https://freesound.org/>.
- [3] [n. d.]. GIS Data for the 2012 National Insect & Disease Risk Map (NIDRM) Report. <https://www.fs.usda.gov/foresthealth/technology>
- [4] [n. d.]. iNaturalist, <https://www.inaturalist.org>.
- [5] [n. d.]. Radio aporee: Maps - sounds of the world, <https://aporee.org>.
- [6] Luca Maria Aiello, Rossano Schifanella, Daniele Quercia, and Francesco Aletta. 2016. Chatty maps: constructing sound maps of urban areas from social media data. *Royal Society open science* 3, 3 (2016), 150690.
- [7] Alex Alemi, Ian Fischer, Josh Dillon, and Kevin Murphy. 2017. Deep Variational Information Bottleneck. In *ICLR*. <https://arxiv.org/abs/1612.00410>

- [8] Rahaf Aljundi, Yash Patel, Milan Sulc, Nikolay Chumerin, and Daniel Olmeda Reino. 2023. Contrastive classification and representation learning with probabilistic interpretation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 6675–6683.
- [9] Benjamin Brodie, Subash Khanal, Muhammad Usman Rafique, Connor Greenwell, and Nathan Jacobs. 2021. Hierarchical Probabilistic Embeddings for Multi-View Image Classification. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. 1011–1014. <https://doi.org/10.1109/IGARSS47720.2021.9554405>
- [10] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. 2020. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3884–3892.
- [11] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Owen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. *arXiv preprint arXiv:2311.07919* (2023).
- [12] Sanghyuk Chun. 2024. Improved Probabilistic Image-Text Representations. In *International Conference on Learning Representations (ICLR)*.
- [13] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8415–8424.
- [14] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. 2022. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. In *Advances in Neural Information Processing Systems*. Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). <https://openreview.net/forum?id=WBhqpF6KYH>
- [15] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [16] Margaret Sibylle Engel, André Fiebig, Carmella Pfaffenbach, and Janina Fels. 2021. A review of the use of psychoacoustic indicators on soundscape studies. *Current Pollution Reports* (2021), 1–20.
- [17] EROS. [n. d.]. National Land Cover Database. <https://www.usgs.gov/centers/eros/science/national-land-cover-database>
- [18] International Organization for Standardization. 2014. ISO 12913-1: 2014 acoustics—Soundscape—part 1: definition and conceptual framework. ISO, Geneva (2014).
- [19] David Montes González, Juan Miguel Barrigón Morillas, and Guillermo Rey-Gozalo. 2023. Effects of noise on pedestrians in urban environments where road traffic is the main source of sound. *Science of the total environment* 857 (2023), 159406.
- [20] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 976–980.
- [21] 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.
- [22] Di Hu, Xuhong Li, Lichao Mou, Pu Jin, Dong Chen, Liping Jing, Xiaoxiang Zhu, and Dejing Dou. 2020. Cross-task transfer for geotagged audiovisual aerial scene recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV* 16. Springer, 68–84.
- [23] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. 2021. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems* 34 (2021), 29406–29419.
- [24] Vladimir Iashin and Esa Rahtu. 2021. Taming Visually Guided Sound Generation. In *British Machine Vision Conference (BMVC)*.
- [25] Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaxing Zhang, Tetsuya Sakai, and Yujiu Yang. 2023. MAP: Multimodal Uncertainty-Aware Vision-Language Pre-Training Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23262–23271.
- [26] Subash Khanal, Srikanth Sastry, Ayush Dhakal, and Nathan Jacobs. 2023. Learning Tri-modal Embeddings for Zero-Shot Soundscape Mapping. In *British Machine Vision Conference (BMVC)*.
- [27] Peter Lercher and Angel M Dzhambov. 2023. Soundscape and Health. In *Soundscape: Humans and Their Acoustic Environment*. Springer, 243–276.
- [28] Matteo Lionello, Francesco Aletta, and Jian Kang. 2020. A systematic review of prediction models for the experience of urban soundscapes. *Applied Acoustics* 170 (2020), 107479.
- [29] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*. 638–647.
- [30] Efstathios Margaritis and Jian Kang. 2017. Soundscape mapping in environmental noise management and urban planning: case studies in two UK cities. *Noise mapping* 4, 1 (2017), 87–103.
- [31] Andrei Neculai, Yanbei Chen, and Zeynep Akata. 2022. Probabilistic compositional embeddings for multimodal image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4547–4557.
- [32] Kenneth Ooi, Zhen-Ting Ong, Karn N Watcharasupat, Bhan Lam, Joo Young Hong, and Woon-Seng Gan. 2023. ARAUS: A large-scale dataset and baseline models of affective responses to augmented urban soundscapes. *IEEE Transactions on Affective Computing* (2023).
- [33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
- [34] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. 2016. Ambient sound provides supervision for visual learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I* 14. Springer, 801–816.
- [35] Judicaël Picaut, Nicolas Fortin, Erwan Bocher, Gwendall Petit, Pierre Aumond, and Gwenaël Guillaume. 2019. An open-science crowdsourcing approach for producing community noise maps using smartphones. *Building and Environment* 148 (2019), 20–33.
- [36] 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>
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [38] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candito, Matt Uyttendaele, and Trevor Darrell. 2023. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4088–4099.
- [39] Tawfiq Salem, Menghua Zhai, Scott Workman, and Nathan Jacobs. 2018. A multimodal approach to mapping soundscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2524–2527.
- [40] Yukun Su, Guosheng Lin, Ruizhou Sun, Yun Hao, and Qingyao Wu. 2021. Modeling the uncertainty for self-supervised 3d skeleton action representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 769–778.
- [41] Li Tao, Xueteng Wang, and Toshihiko Yamasaki. 2020. Self-supervised video representation learning using inter-intra contrastive framework. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2193–2201.
- [42] 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>
- [43] Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. 2023. ProbLVM: Probabilistic Adapter for Frozen Vision-Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1899–1910.
- [44] Junsheng Wang, Tiantian Gong, Zhixiong Zeng, Changchang Sun, and Yan Yan. 2022. C3CMR: Cross-Modality Cross-Instance Contrastive Learning for Cross-Media Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) (MM '22). Association for Computing Machinery, New York, NY, USA, 4300–4308. <https://doi.org/10.1145/3503161.3548263>
- [45] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4563–4567.
- [46] Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP.
- [47] Donghuo Zeng, Jianming Wu, Gen Hattori, Rong Xu, and Yi Yu. 2023. Learning explicit and implicit dual common subspaces for audio-visual cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 2s (2023), 1–23.
- [48] Tianhong Zhao, Xiucheng Liang, Wei Tu, Zhengdong Huang, and Filip Biljecki. 2023. Sensing urban soundscapes from street view imagery. *Computers, Environment and Urban Systems* 99 (2023), 101915.

A DATASET CREATION

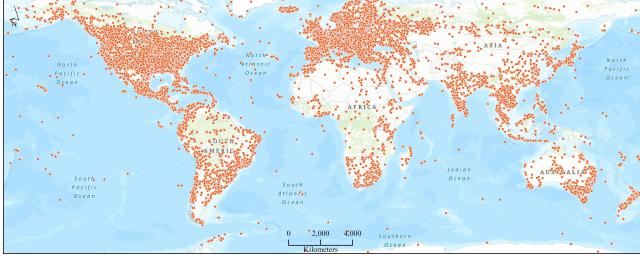


Figure 5: Distribution of samples in the *GeoSound* dataset.

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 [21] 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* [4], *YFCC100M* [42], *Radio Aporee* [5], 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 5 illustrates the geospatial distribution of data samples in the *GeoSound* dataset worldwide.

A.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 [21], 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 5: Distribution of *GeoSound* Dataset Across Splits and Audio Sources.

A.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 5. Our data split strategy on *GeoSound* dataset is described as follows:

- (1) We divide the world into $1^\circ \times 1^\circ$ 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.

B 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 6, we present sets of samples with high uncertainty and low uncertainty for *Bing* satellite imagery in our test set.



Figure 6: 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.

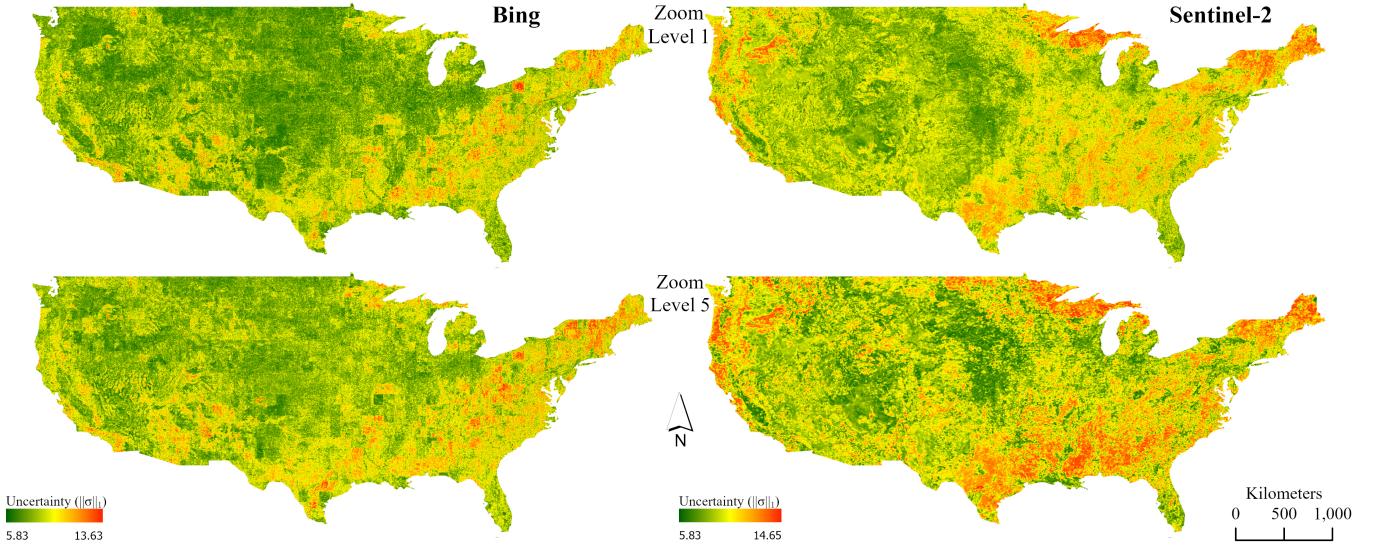


Figure 7: Uncertainty map of the satellite image embeddings for the USA. Uncertainty at each location is approximated as the $\|\sigma\|_1$ of the probabilistic embeddings obtained from our best-performing model trained with *Bing* and *Sentinel-2* imagery, respectively, without any metadata.

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 7. 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

meta/train	text/eval	meta/eval	R@10%	MdR
GeoSound-Bing				
✗	✗	✗	0.423	1401
✓	✗	✗	0.425	1359
✓	✗	✓	0.828	261
✓	✓	✗	0.776	213
✓	✓	✓	0.901	113
GeoSound-Sentinel2				
✗	✗	✗	0.474	1101
✓	✗	✗	<u>0.464</u>	<u>1144</u>
✓	✗	✓	0.802	294
✓	✓	✗	0.730	256
✓	✓	✓	0.872	142
SoundingEarth				
✗	✗	✗	0.514	547
✓	✗	✗	<u>0.514</u>	<u>550</u>
✓	✗	✓	0.563	454
✓	✓	✗	<u>0.651</u>	<u>315</u>
✓	✓	✓	0.690	264

Table 6: Evaluation of PSM under different settings for the datasets: GeoSound-Bing, GeoSound-Sentinel2, and SoundingEarth. The results are Recall@10% (R@10%) and Median-Rank (MdR) for Image-to-Audio retrieval with satellite image at zoom level 1. Results highlighted with underlined text represent the evaluation of PSM without using any metadata during inference, for comparison with other evaluations presented in the main paper.

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.

C METADATA DEPENDENCY

We provide a comparison of our framework’s performance without using metadata during evaluation, as shown in Table 6. As observed, the performance of PSM with text added to query embeddings, without including any other metadata, remains comparable to the best setting, which includes both text and metadata during retrieval. Using metadata without text in the query improves performance compared to not using either. This highlights the importance of encoding the dynamic nature of the soundscape based on location (latitude and longitude), season (month), and hour of the day (time). It is worth noting that, as shown in Table 4 of the main paper, which presents an ablation study of different metadata components, *audio source* is the most important metadata. Therefore, during inference, the metadata type *audio source* can be inferred by understanding the unique nature of sounds from each platform of audio data used to curate the dataset, as described in Section A of this supplemental material.

D SOUNDSCAPE MAPS

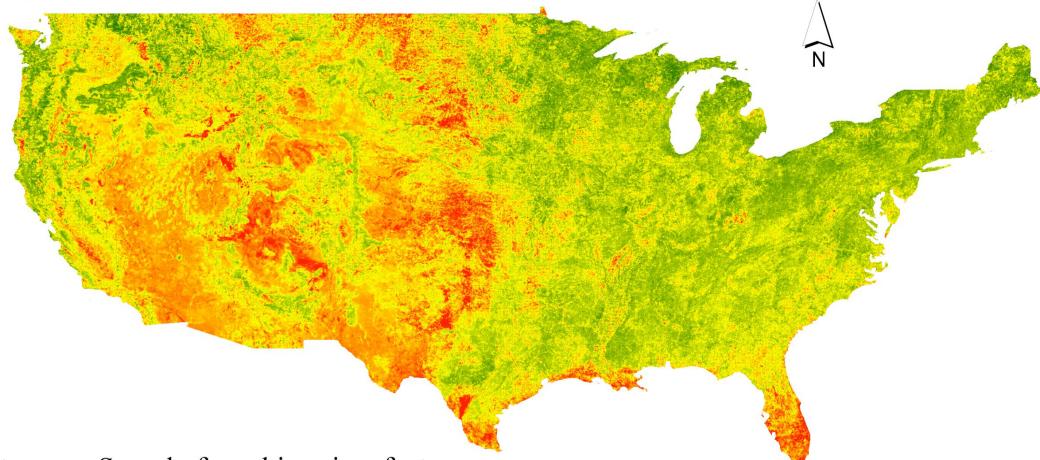
In Figure 8, we present examples of country-scale soundscape maps over the USA. These maps were generated using our best-performing model trained on *Sentinel-2* imagery without any metadata. In this demonstration, we utilize *Sentinel-2* imagery covering the USA at zoom-level 1. In the figure, for the 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 9, 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 [36] 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 9, each soundscape map is accompanied by a land cover map [17] of the respective region for reference.

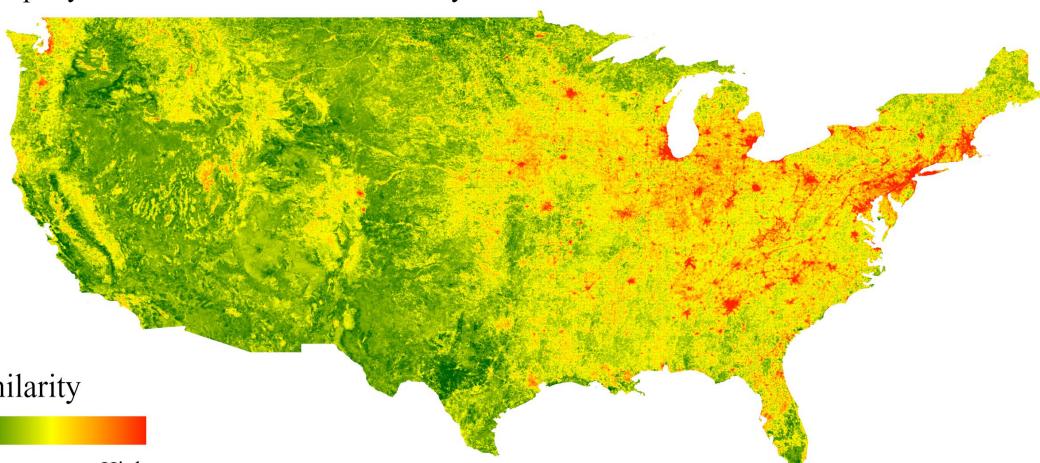
As observed in Figure 9, geographic regions expected to be related to the query demonstrate high similarity scores. For example, for the audio query described by the audio class *cow*, we can see that urban regions around cities like *Memphis* and *Toledo* have low similarity scores, while rural areas (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 9, 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 phenomenon highlights the inherent trade-off in spatial resolution when using different zoom levels for soundscape mapping. 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).

Text query: Sound of animals on a farm



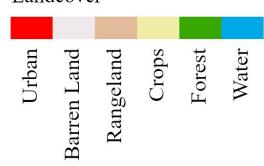
Text query: Sound of machines in a factory



Similarity



Landcover



Kilometers
0 500 1,000

Figure 8: Two soundscape maps of the continental United States, generated from *Sentinel-2* image embeddings, accompanied by a land cover map for reference [17].

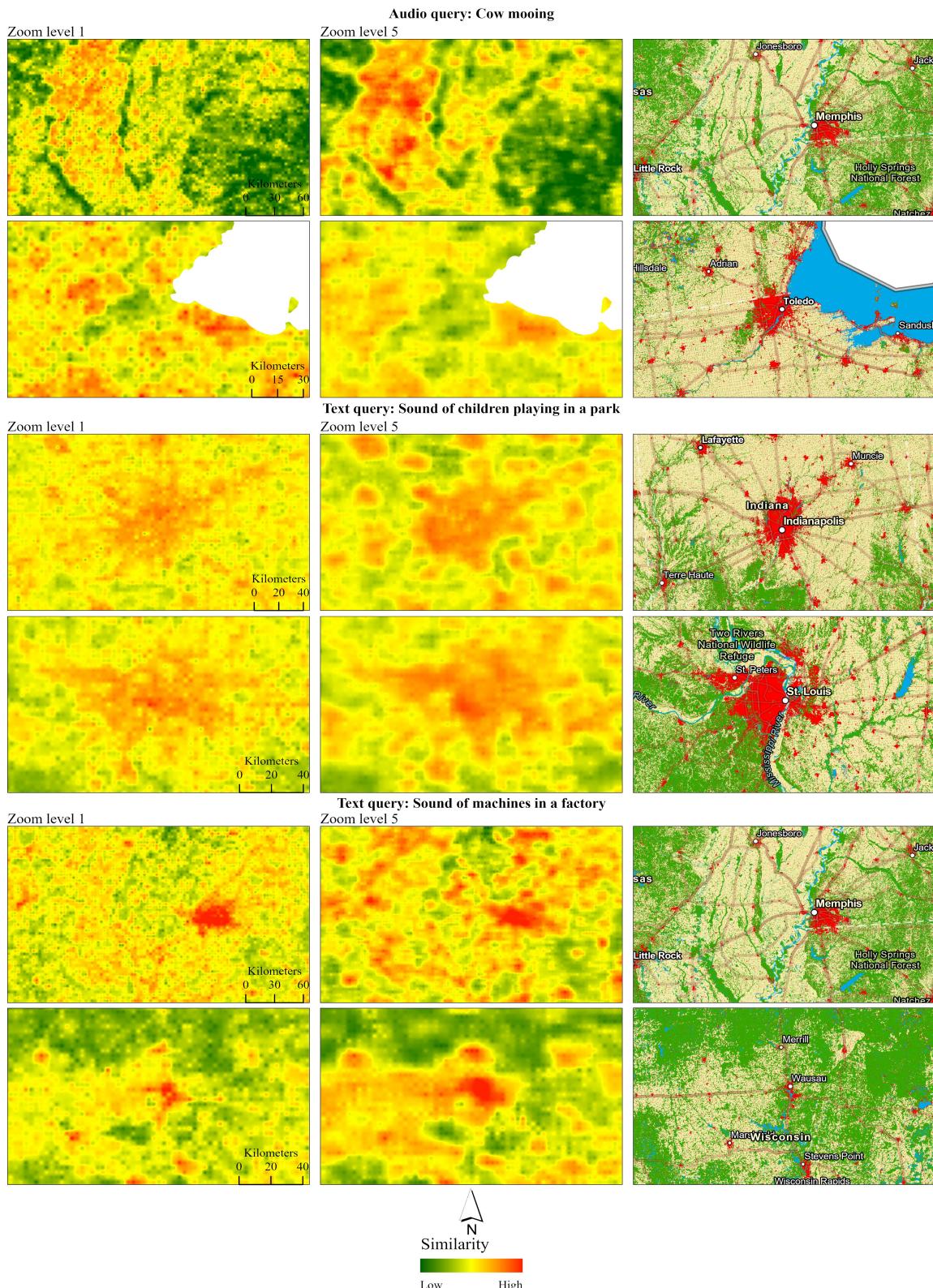


Figure 9: Soundscape maps over smaller geographic areas, computed using similarity scores between respective queries and embeddings from *Sentinel-2* satellite imagery at two zoom levels.