Learning Tri-modal Embeddings for Zero-Shot Soundscape Mapping

Subash Khanal
k.subash@wustl.edu
Srikumar Sastry
s.sastry@wustl.edu
Aayush Dhakal
a.dhakal@wustl.edu
Nathan Jacobs
jacobsn@wustl.edu

Computer Science & Engineering
Washington University in St. Louis
St. Louis, MO, USA

Abstract

We focus on the task of soundscape mapping, which involves predicting the most probable sounds that could be perceived at a particular geographic location. We utilise recent state-of-the-art models to encode geotagged audio, a textual description of the audio, and an overhead image of its capture location using contrastive pre-training. The end result is a shared embedding space for the three modalities, which enables the construction of soundscape maps for any geographic region from textual or audio queries. Using the SoundingEarth dataset, we find that our approach significantly outperforms the existing SOTA, with an improvement of image-to-audio Recall@100 from 0.256 to 0.450. Our code is available at https://github.com/mvrl/geoclap.

1 Introduction

Sound is one of the fundamental senses that helps us reason about our environment. There exists an intricate relationship between the visual appearance and sound of a location [15, 16]. Learning about the type of sound at a geographic location allows one to understand many high-level concepts of the area. For example, just by hearing the sound of traffic, we can imagine the location to be an urban setting with a rush of cars and people, whereas the sound of sea waves might elicit the beautiful scenery of a beach.

There have been several studies conducted on different cities around the world attempting to understand human perception of various types of environmental sound [1, 3, 15, 22, 28, 30]. Moreover, it has been established that there is a strong correlation between the physiological and psychological health of a person and the environmental sound condition they live in [8, 21, 33]. Therefore, understanding the soundscape for a given geographic area can be of great importance to policymakers focused on urban planning and environmental noise management. Soundscapes also serve value to the general public for whom environmental sound plays a vital role in decisions such as buying a house or setting up a business.

Most of the existing works on creating soundscape focus on crowd-sourcing human perception of sound in their surroundings [1, 3, 22, 30, 40]. While serving as an important tool for understanding the sound distribution of a region, such approaches have two major limitations. First, the abstraction of sound into a fixed set of indicators and psycho-acoustic descriptors limits our ability to have a complete picture of underlying physical factors associated with sound. Second, such soundscapes are usually created for only highly visited places in the world, creating massive sparsity of soundscapes on a global scale. In order to solve both of these limitations, we propose to leverage the intrinsic relationship between sound and visual cues of the location and learn to directly predict the most probable sound that could be heard at any given location. Specifically, we train a multi-modal deep learning framework that learns a shared embedding space where the sound that is most likely to come from a given location, is pulled closer while pushing other unlikely sounds farther apart. We represent the location (latitude, longitude) by an overhead image of size $H \times W$ centered around it. Once trained, our multi-modal embedding space and free availability of overhead imagery makes it possible for us to create soundscape maps for any area in the world.

One of the successful approaches to learning shared embedding space between different modalities is contrastive learning. In recent years, contrastive learning between image and text [32]; image, text, and audio [17]; text, audio [9, 12, 38]; overhead image and audio [19] has been an effective self-supervised training objective to learn a multi-modal embedding space. Such a space has an understanding of the correspondence between the modalities that can be transferred to various downstream tasks, where impressive results have been observed. Motivated by these works, we also adopt contrastive learning as our pre-training strategy to learn a multi-modal embedding space. However, unlike the prior works, we are interested in incorporating geographic knowledge into the embedding space learned by audio-language pre-training. We achieve this by adding an overhead image, capturing the geographic context of a scene, as an additional modality in our contrastive learning framework. With the shared embedding space that has knowledge of correspondence between audio and its corresponding overhead image, we can then formulate the task of soundscape mapping as a cross-modal retrieval problem, where the objective is to predict the most likely sound from a gallery of N sounds given an overhead image.

Our work builds upon a prior work [19] that introduced the *SoundingEarth* dataset containing over $50k$ geotagged audios paired with their corresponding overhead image. The objective of work by Heidler *et al*. [19] was to learn a good audio-visual embedding space useful to be transferred for different downstream tasks in remote sensing. However, in the interest of learning an embedding space to create accurate soundscapes, our work is focused on improving the task of cross-modal retrieval. In this regard, we utilise weights of publicly available modality-specific SOTA models. Moreover, unlike Heidler *et al*., who build an embedding space capturing two modalities (overhead-image and audio), we propose to also incorporate textual description of audio into the embedding space. This essentially creates a tri-modal embedding space with richer understanding of three modalities: overhead-image, audio, and text. We call our framework GeoCLAP: Geography-Aware Contrastive Language Audio Pre-training. As demonstrated by our results adding the textual modality improves the representational capability of both overhead-image and audio encoders. Moreover, with an understanding of three modalities, we are now able to create soundscapes either from a textual or audio query for any geographic region. The main contributions of our work are as follows:

- We significantly improve the prior baseline on the task of cross-modal retrieval of overhead image to sound and vice-versa.

- We build a tri-modal embedding space that has an understanding of overhead image, audio, and textual description of audio at a given location.
- We demonstrate a simple and scalable way of creating soundscape for any geographic area using either a textual or audio query.

2 Related Work

2.1 Soundscape Mapping

The soundscape of a geographic region can be defined as the acoustic environment perceived by individuals within its context [14]. There exists a large body of work focusing on the problem of soundscape mapping [1, 3, 13, 16, 22, 26, 28, 30, 40, 41]. In these works, soundscape mapping is formulated as a framework containing three components: indicators, descriptors, and a predictive model that maps indicators to descriptors. Indicators are psycho-acoustic measures (for example, sound pressure level, loudness, spectral slope, etc.) which determine the perceived value of descriptors (for example, pleasant, unpleasant, eventful, etc.). In this paper, we refer to this line of work as perceptual soundscape mapping.

One of the common findings from the literature of perceptual soundscape mapping is that there exists a strong correlation between the human perception of sound and the environmental variables of the scene such as building, road category, etc. [15]. Utilising this correlation between sound and visual cues, there have been a few works that use deep learning to learn a shared embedding space between sound and either ground level image [29] or overhead image [20] of the scene. This multimodal learning approach leads to improved performance on visual tasks such as aerial scene recognition [20], image classification [29], and object detection [29]. Closer to our work, a few prior works [6, 25, 27, 39] focus on the task of cross-modal image-to-voice retrieval. Such tasks require a dataset containing overhead imagery paired with spoken audio captions, which is very limited. Moreover, instead of learning from speech, we are interested in learning from free-form audio such as field recordings, natural sounds, etc. which capture diverse concepts of the location. Another closer work by Salem *et al*. [35], proposed learning a shared embedding space between audio, overhead image, and ground level image, enabling them to predict a distribution over sound clusters from an overhead image. The problem formulation of soundscape mapping in our work is similar to [35]. However, the striking difference as well as the strength of our work is that leveraging the power of contrastive language audio pre-training (CLAP), we are able to create soundscape conditioned on any textual or audio query. In doing so, we still retain the ability to create soundscape with desired set of sound categories in a zero-shot manner.

2.2 Contrastive Learning

Radford *et al*., in their seminal work, CLIP [32], trained large image-text dataset using contrastive loss and demonstrated it's impressive zero-shot performance on many computer vision tasks. AudioCLIP [17], extends CLIP to three modalities: image, text, and audio. Such tri-modal embedding space enables one to perform query between three pairs of modalities. Wav2clip [37], distilled the knowledge of CLIP embedding space by freezing the image encoder of CLIP and contrastively training an audio encoder to learn a new embedding space shared by audio and a corresponding image. With similar training objective as CLIP, an-

other work CLAP [12] performs contrastive learning between audio and natural language. CLAP training has proven to be an effective strategy with impressive audio retrieval performance [9]. Inspired by this, Wu *et al*. [38] further improved the CLAP's performance by training on large-scale data with effective audio feature fusion and text augmentation strategies. We refer the work by Wu *et al*. [38] as L-CLAP in our paper and use the pre-trained encoders from L-CLAP to embed audio and text for GeoCLAP pre-training.

Our work takes motivation from the proven performance of contrastive learning as an effective pre-training strategy. The focus of our work is soundscape mapping. The embedding space for such tasks should have an understanding of geography of a location where the sound is coming from [4]. Therefore, we propose to learn an embedding space trained contrastively on three modalities: overhead image, text, and audio.

2.3 Pretrained Models

Availability of modality specific pre-trained models trained with various self-supervision objectives have proven to be crucial in bringing performance improvement in various tasks in remote sensing [36]. In the recent years, masked auto-encoders (MAE) [18] based models trained on satellite imagery have demonstrated to be a good starting checkpoints to be fine-tuned for various downstream tasks [7, 34]. In our work, we start with the pre-trained weights of Vision Transformer (ViT) [11] encoder of SATMAE [7] as the overhead-image encoder for GeoCLAP. SATMAE [7] was pre-trained on large-scale (over 700K) satellite imagery of the world. To learn representations for audio and text, we use L-CLAP's pre-trained encoders. It uses HTSAT [5] as the audio encoder and RoBERTa [23] as the text encoder. HTSAT is a swin-transformer [24] based model with SOTA performance on various audio classification tasks. RoBERTa is a powerful transformer-based language model trained with improved design choices than BERT [10]. L-CLAP [38] was contrastively pre-trained on over 630K audio-text paired dataset.

3 Approach

We present a detailed description of our approach, including the high-level problem formulation, a description of our primary evaluation dataset, and a detailed description of the network architecture and training procedure for our method, GeoCLAP.

3.1 Problem formulation

The objective of our work is to learn a shared embedding space that allows us to predict the most probable sounds that can be heard at a given geographic location. This can be represented as $s^* = \max_s P(s|l)$ where $P(s|l)$ represents the conditional distribution of sounds for a given location l and s^* is the most likely sound. Unfortunately, direct conditioning on location does not generalize to regions without a large number of training samples, which means truly global mapping wouldn't be possible. On the other hand, overhead imagery has a strong correlation to the type of sound at a given location and is freely available across the globe. Therefore, in our work, we represent the location indirectly, using an overhead image $I(l)$ of the location. We learn a conditional distribution $P(s|I(l))$, which is able to make high-resolution predictions even for regions without training samples.

3.2 Dataset

We use the *SoundingEarth* dataset to train and evaluate our method. The dataset contains more than 50k geotagged audio recordings from 136 countries and overhead image pairs. The overhead images have size of 1024×1024 collected from *Google Earth* with an approximate ground-sample distance (GSD) of 0.2 meters (m). Audio data in the dataset was collected from the project *Radio Aporee:::Maps* [7], which hosts an online platform dedicated to creating a global soundmap. It contains diverse audio recordings from urban, rural and natural environments, published under the creative commons license. For our project, we remove the audio files with a sampling frequency less than 16k. This yields a dataset size of 50 792 samples.

The high-resolution *Google Earth* imagery is not available to be used freely. Therefore, in order to have the ability to globally scale soundscape mapping, we augment the existing *SoundingEarth* dataset by including freely available lower-resolution images. Specifically, we use the RGB bands of the *Sentinel-2 cloudless* imagery with 10m *GSD*. For each location, we download a 256×256 image tile with the coverage radius of 512m centered at that location.

3.3 GeoCLAP

Figure 1 represents the overall framework of GeoCLAP. Given a geotagged audio X_k^a, textual description of the audio X_k^t, and an overhead image at a given location X_k^i, where (X_k^a, X_k^t, X_k^i) is one audio-text-image triplet. We obtain embeddings for each modalities by passing through modality-specific encoder and linear projection layer, yielding embeddings with the same dimension for audio, text, and overhead image, respectively.

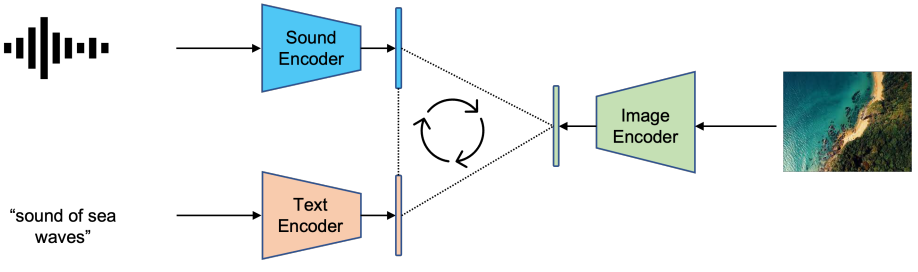

Figure 1: GeoCLAP: A tri-modal contrastive learning framework to learn shared embedding space between overhead image, sound, and textual description of the corresponding sound.

$$E_k^a = g_{audio}(f_{audio}(X_k^a)) \tag{1}$$

$$E_k^t = g_{text}(f_{text}(X_k^t)) \tag{2}$$

$$E_k^i = g_{image}(f_{image}(X_k^i)) \tag{3}$$

where (f_{audio}, g_{audio}), (f_{text}, g_{text}), (f_{image}, g_{image}) are (encoder, linear projection layer) pairs producing $l2$-normalized d dimensional embeddings: E_k^a, E_k^t, and E_k^i, for audio, text, and overhead image respectively.

GeoCLAP is trained on embedding triplets using contrastive learning objective similar to CLIP [32] for all three pairs of embeddings:

$$L_{at} = \frac{1}{2N}\sum_{k=1}^{N}\left(log\frac{exp(E_k^a.E_k^t/\tau_{at})}{\sum_{j=1}^{N}exp(E_k^a.E_j^t/\tau_{at})} + log\frac{exp(E_k^t.E_k^a/\tau_{at})}{\sum_{j=1}^{N}exp(E_k^t.E_j^a/\tau_{at})}\right) \tag{4}$$

$$L_{ai} = \frac{1}{2N}\sum_{k=1}^{N}\left(log\frac{exp(E_k^a.E_k^i/\tau_{ai})}{\sum_{j=1}^{N}exp(E_k^a.E_j^i/\tau_{ai})} + log\frac{exp(E_k^i.E_k^a/\tau_{ai})}{\sum_{j=1}^{N}exp(E_k^i.E_j^a/\tau_{ai})}\right) \tag{5}$$

$$L_{ti} = \frac{1}{2N}\sum_{k=1}^{N}\left(log\frac{exp(E_k^t.E_k^i/\tau_{ti})}{\sum_{j=1}^{N}exp(E_k^t.E_j^i/\tau_{ti})} + log\frac{exp(E_k^i.E_k^t/\tau_{ti})}{\sum_{j=1}^{N}exp(E_k^i.E_j^t/\tau_{ti})}\right) \tag{6}$$

where, N is the training batch size and τ_{at}, τ_{ai}, and τ_{ti} are learnable temperature parameters used to scale logits in loss computation for each pairs of embeddings.

Combining equations 4, 5, and 6, the final loss for which GeoCLAP is trained is as follows:

$$L = L_{at} + L_{ai} + L_{ti} \tag{7}$$

4 Experimental Details

4.1 Data Preprocessing

For audio preprocessing, we convert each audio sample into mel-spectrogram using the default settings: `{feature_size=64, sampling_rate=48000, hop_length=480, max_length_s=10, fft_window_size=1024}` provided in the `HuggingFace`-wrapper:`ClapProcessor` for the pre-trained L-CLAP model `clap-htsat-fused`.

In the *SoundingEarth* dataset, most of the audio recordings (except 6333 samples) are also accompanied by a brief description and a title uploaded by the contributor. In order to have textual description for all audio recordings as well as to further encode geographic information in text, we use a python client, `geopy` to obtain the address of the location and append an additional sentence, *"The location of the sound is:{address}."* to the textual description of each sample. For example, for the geolocation $(52.509663, 13.376481)$, the added sentence would be *"The location of the sound is: Potsdamer Platz, Tiergarten, Mitte, Berlin, 10785, Germany"*. Following L-CLAP, we use `RobertaTokenizer` with the parameter `max_length` set to 77.

For overhead imagery, we adopt the same data augmentation as SATMAE [7]. We perform *RandomResizedCrop* with parameters:`{input_size=224, scale=(0.2,1.0)}`, followed by a *RandomHorizontalFlip*, during training. During inference, we extract a 224×224 center crop of the image.

4.2 Implementation and metrics

We implement our code in `Pytorch` and utilise `HuggingFace` for loading L-CLAP encoders and their respective data pre-processing wrappers. We split the dataset with ratio 70:10:20 yielding a total of 35 554, 5 079, and 10 159 samples into training, validation, and test split, respectively. For experiments regarding the baseline, we ran the publicly available code for [19] using the data splits of our study. We used the experimental setting for their best reported results on cross-modal retrieval task, which is as follows:`{batch_size=256 encoders=ResNet18, latent_dim=128, loss=SymmetricCL, tau=0.2}`. The baseline was trained for 300 `epochs` with `Adam` optimizer and learning rate of $1e-3$.

4.2.1 Encoders

We use the pre-trained model `clap-htsat-fused` [38] to encode audio and text. The audio encoder used in our study, HTSAT, has 4 swin-transformer blocks with hidden feature dimension of 768. The text encoder RoBERTa from [38] used in our study, has 12 transformer blocks with hidden feature dimension of 768. For both audio and text encoders, we take the output of their respective L-CLAP's projection layer producing 512-dimensional embeddings. For encoding overhead image, we use the pre-trained `vit_base_patch16` encoder of SATMAE [7]. It processes input as a sequence of 16×16 image patches passing through 12 layers of transformer blocks. In order to match dimension with audio and text embeddings, we pass the output from SATMAE encoder to a ReLU activation followed by a 512-dimension linear layer. Starting from weights of these pre-trained encoders, we conduct two set of experiments. First, we allow only the overhead-image encoder to train while freezing L-CLAP. Second, we allow fine-tuning of all encoders in our framework.

4.2.2 Training

We train GeoCLAP using the contrastive loss objective presented in Equation 7. We initialize all three learnable temperature parameters to 0.07. We also run experiments with and without using *text* in our framework. While using text, we further experiment the impact of adding an additional sentence describing detailed address of the location to the text. For experiments where we use overhead image and audio only, we train our model with image-audio contrastive loss represented by Equation 5. Moreover, for experiments using overhead image, audio, and text, while keeping the L-CLAP encoders frozen, we train with $Loss = L_{ai} + L_{ti}$. We use a training batch size of 256 for the baseline, and our experiments with frozen L-CLAP, while using batch size of 128 for experiments allowing fine tuning of L-CLAP. We use the `Adam` optimizer and set the initial learning rate to $5e-5$. We use `weight_decay=0.2` and `betas=(0.9,0.98)`. We use cosine annealing learning rate scheduler with number of warm up iterations set to 2000. We set `max_epochs` to 100 for experiments with frozen L-CLAP and 30 for experiments allowing fine tuning of L-CLAP.

4.2.3 Metrics

Following Heidler *et al.* [19], we use Recall@100 and Median Rank (Median-R) of the ground-truth as the evaluation metrics of our approach. We use the test set containing 10 159 samples as the gallery for both image-to-sound and sound-to-image retrieval evaluation.

5 Evaluation

5.1 Experiments with SoundingEarth data

Table 1 shows the results of our experiments with the *SoundingEarth* dataset while using the original overhead imagery of $0.2m$ resolution. One of the interesting results from this table is that by just using frozen pre-trained audio encoder from L-CLAP [38], while allowing only overhead-image encoder to be fine-tuned, we already get about 10 points improvement in cross-modal retrieval. This highlights the advantage of leveraging rich representation space of pre-trained models like L-CLAP. However, when we introduce text modality into training, while still keeping both text and audio encoders frozen, the image-to-sound Recall@100

Method					Image2Sound		Sound2Image	
Experiment	Image Encoder	Text-Audio Encoder	Text	Address	R@100	Median-R	R@100	Median-R
Baseline [13]	ResNet18	ResNet18	✗	✗	0.256	814	0.250	816
ours	SATMAE	L-CLAP-frozen	✗	✗	0.352	360	0.348	369
ours	SATMAE	L-CLAP-frozen	✓	✗	0.328	428	0.325	428
ours	SATMAE	L-CLAP-frozen	✗	✓	0.298	546	0.295	544
ours	SATMAE	L-CLAP-frozen	✓	✓	0.317	439	0.311	443
ours	SATMAE	L-CLAP	✗	✗	0.384	230	0.385	237
ours	SATMAE	L-CLAP	✓	✗	0.423	172	0.419	175
ours	SATMAE	L-CLAP	✗	✓	0.432	166	0.431	167
ours	SATMAE	L-CLAP	✓	✓	**0.434**	**159**	**0.434**	**167**

Table 1: Cross-modal retrieval performance for models using 0.2m GSD overhead imagery.

drops to 0.32. L-CLAP was trained on large corpus of text-audio pairs where textual description of audio have relatively high quality. However, the primary focus of the *SoundingEarth* dataset has been to collect geotagged audio from all around the world and associate them with high-resolution overhead imagery. We observed that the textual descriptions of audio in the *SoundingEarth* dataset are noisy and do not reflect the type of textual prompts L-CLAP models were trained on. In our experiments, we use three different types of texts: textual description of audio, only address of the audio, and text containing both description and address of the audio. We observed that for any type of text, learning with frozen representation lowers the performance when compared to learning with frozen representation of audio alone. With this observation, we decided to allow fine-tuning of L-CLAP encoders. Accordingly, the performance of our approach noticeably improves to image-to-sound Recall@100 of 0.384 while learning with overhead image and audio. The performance further improves to Recall@100 of 0.423 with Median Rank of 172 when we learn with overhead image, audio, and text. This performance is further improved to Recall@100 of 0.434 with Median Rank of 159, when we add address of the audio location in the text. This is an absolute improvement of the baseline performance by 0.178 points in image-to-sound Recall@100 and 655 in Median Rank. We see similar trends on sound-to-image retrieval task.

5.2 Experiments with Sentinel data

Table 2 shows the results of our experiments with *Sentinel-2 cloudless* imagery with 10*m GSD*. We found that performance in all of our experiments noticeably improved while using lower-resolution overhead imagery. This choice brought in 12.89% relative improvement in the baseline Recall@100 performance as well. We believe the reason for this improvement is the larger coverage of geographic area in a single overhead image with 10m *GSD*. Moreover, the lower-resolution sentinel imagery is inherently blurry offering some regularization effect during training, leading to improved generalizability of our models. Following similar trends as in Table 1, an absolute Recall@100 improvement of about 10 points is observed, when using a pre-trained frozen audio encoder from L-CLAP. Similarly, the retrieval performance improves to 0.396 when the audio encoder is allowed to be fine-tuned. We also observe gain in performance of fine-tuned GeoCLAP models trained with text containing address. The best performance for GeoCLAP trained with all three modalities, yields (Recall@100, Median Rank) of (0.450,143) and (0.447,144) for image-to-sound and sound-to-image retrieval, respectively. Compared to the baseline, this is a relative gain of 55.71% and 57.95% for Recall@100 on tasks: image-to-sound and sound-to-image retrieval, respectively.

Method					Image2Sound		Sound2Image	
Experiment	Image Encoder	Text-Audio Encoder	Text	Address	R@100	Median-R	R@100	Median-R
Baseline [19]	ResNet18	ResNet18	✗	✗	0.289	620	0.283	635
ours	SATMAE	L-CLAP-frozen	✗	✗	0.384	274	0.381	271
ours	SATMAE	L-CLAP-frozen	✓	✗	0.340	369	0.338	367
ours	SATMAE	L-CLAP-frozen	✗	✓	0.311	453	0.304	461
ours	SATMAE	L-CLAP-frozen	✓	✓	0.337	378	0.331	370
ours	SATMAE	L-CLAP	✗	✗	0.396	199	0.396	205
ours	SATMAE	L-CLAP	✓	✗	0.441	152	0.441	155
ours	SATMAE	L-CLAP	✗	✓	0.441	153	0.440	156
ours	SATMAE	L-CLAP	✓	✓	**0.450**	**143**	**0.447**	**144**

Table 2: Cross-modal retrieval performance for models using 10m GSD overhead imagery.

5.3 Zero-Shot Soundscape Mapping

Utilising the rich representation space of our best-performing GeoCLAP model, we demonstrate zero-shot soundscape mapping using both text and audio queries. Soundscape maps, in our work, are the similarity-score heatmaps for a given query. Specifically, we use the appropriate encoder from GeoCLAP to produce an embedding of the query and embeddings for a dense set of overhead images in the region of interest. Then, the cosine similarity score between the query embedding and all overhead image embeddings is overlaid on the corresponding region to yield a soundscape map (Figure 2). In Figure 3, we demonstrate a country-scale soundscape map for the Netherlands. For this, we compute soundscape for three prompts: {*This is a sound of car horn; This is a sound of chirping birds; This is a sound of animal farm*} and overlay them together to create a composite pseudo-color map. We compare this soundscape with ESRI's *Sentinel-2 land cover* classes. We observe a strikingly high correlation between the related land-cover classes with the category of sound likely to be heard at the location. More such soundscape maps can be found in the supplemental material of this paper.

(a) (b)

Figure 2: Soundscape maps along with reference overhead image for two regions. Soundscape created for queries: (a) A textual prompt: *This is a sound of sea waves*; (b) randomly selected sound from the class `chirping_birds` from ESC50 database [31] (green: more probable, white: less probable).

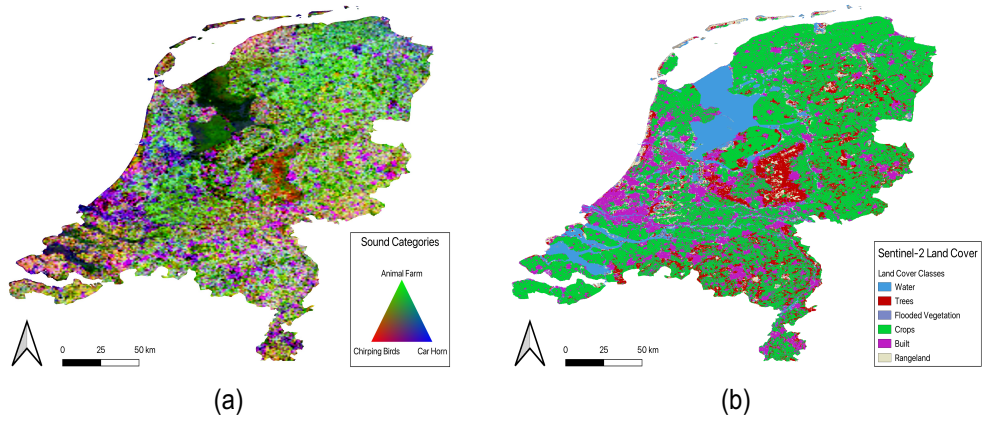

Figure 3: Comparison of (a) Soundscape map of the Netherlands with (b) *Sentinel-2 land cover classes*. The soundscape map was created by querying GeoCLAP with textual prompts for three sound categories: *car horn*, *chirping birds*, and *animal farm*.

6 Conclusion

We proposed GeoCLAP, a contrastive-learning framework capable of embedding the modalities of overhead imagery, audio, and text into a common space. Our approach significantly improves the state of the art for cross-modal retrieval between overhead imagery and audio. We utilise the learned, multi-modal representation space for soundscape mapping, demonstrating a simple and scalable way to create soundscape maps for any geographic area using only satellite imagery and audio or textual queries. With this approach, we can construct global, high-resolution soundmaps with minimal effort.

References

[1] Luca Maria Aiello, Rossano Schifanella, Daniele Quercia, and Francesco Aletta. Chatty maps: constructing sound maps of urban areas from social media data. *Royal Society open science*, 3(3):150690, 2016.

[2] Radio Aporee. https://aporee.org/maps.

[3] Eiji Aramaki and Shoko Wakamiya. Image and sound of the city. In *The Social City: Space as Collaborative Media to Enhance the Value of the City*, pages 205–214. Springer, 2023.

[4] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[5] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classifica-

tion and detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[6] Yaxiong Chen, Xiaoqiang Lu, and Shuai Wang. Deep cross-modal image–voice retrieval in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 58 (10):7049–7061, 2020.

[7] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

[8] Peng Cui, Tingting Li, Zhengwei Xia, and Chunyu Dai. Research on the effects of soundscapes on human psychological health in an old community of a cold region. *International Journal of Environmental Research and Public Health*, 19(12):7212, 2022.

[9] Soham Deshmukh, Benjamin Elizalde, and Huaming Wang. Audio retrieval with wavtext5k and clap training. *arXiv preprint arXiv:2209.14275*, 2022.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *Proceedings of the International Conference on Learning Representations*, 2021.

[12] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[13] Margret Sibylle Engel, André Fiebig, Carmella Pfaffenbach, and Janina Fels. A review of the use of psychoacoustic indicators on soundscape studies. *Current Pollution Reports*, pages 1–20, 2021.

[14] International Organization for Standardization. Iso 12913-1: 2014 acoustics—soundscape—part 1: definition and conceptual framework. *ISO, Geneva*, 2014.

[15] Luis Garzón, Luis Bravo-Moncayo, Julián Arellana, and Juan de Dios Ortúzar. On the relationships between auditory and visual factors in a residential environment context: A sem approach. *Frontiers in Psychology*, 14, 2023.

[16] David Montes González, Juan Miguel Barrigón Morillas, and Guillermo Rey-Gozalo. Effects of noise on pedestrians in urban environments where road traffic is the main source of sound. *Science of the total environment*, 857:159406, 2023.

[17] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[19] Konrad Heidler, Lichao Mou, Di Hu, Pu Jin, Guangyao Li, Chuang Gan, Ji-Rong Wen, and Xiao Xiang Zhu. Self-supervised audiovisual representation learning for remote sensing data. *International Journal of Applied Earth Observation and Geoinformation*, 116:103130, 2023.

[20] Di Hu, Xuhong Li, Lichao Mou, Pu Jin, Dong Chen, Liping Jing, Xiaoxiang Zhu, and Dejing Dou. Cross-task transfer for geotagged audiovisual aerial scene recognition. In *Proceedings of the European Conference on Computer Vision*. Springer, 2020.

[21] Peter Lercher and Angel M Dzhambov. Soundscape and health. In *Soundscapes: Humans and Their Acoustic Environment*, pages 243–276. Springer, 2023.

[22] Matteo Lionello, Francesco Aletta, and Jian Kang. A systematic review of prediction models for the experience of urban soundscapes. *Applied Acoustics*, 170:107479, 2020.

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[25] Guo Mao, Yuan Yuan, and Lu Xiaoqiang. Deep cross-modal retrieval for remote sensing image and audio. In *10th IAPR workshop on pattern recognition in remote sensing*, 2018.

[26] Efstathios Margaritis and Jian Kang. Soundscape mapping in environmental noise management and urban planning: case studies in two uk cities. *Noise mapping*, 4(1):87–103, 2017.

[27] Hailong Ning, Bin Zhao, and Yuan Yuan. Semantics-consistent representation learning for remote sensing image–voice retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

[28] Kenneth Ooi, Zhen-Ting Ong, Karn N Watcharasupat, Bhan Lam, Joo Young Hong, and Woon-Seng Gan. Araus: A large-scale dataset and baseline models of affective responses to augmented urban soundscapes. *IEEE Transactions on Affective Computing*, 2023.

[29] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision*, 2016.

[30] Judicaël Picaut, Nicolas Fortin, Erwan Bocher, Gwendall Petit, Pierre Aumond, and Gwenaël Guillaume. An open-science crowdsourcing approach for producing community noise maps using smartphones. *Building and Environment*, 148:20–33, 2019.

[31] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the Association for Computing Machinery Conference on Multimedia*, 2015.

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

[33] Antonella Radicchi, Pınar Cevikayak Yelmi, Andy Chung, Pamela Jordan, Sharon Stewart, Aggelos Tsaligopoulos, Lindsay McCunn, and Marcus Grant. Sound and the healthy city. *Cities & Health*, 5(1-2):1–13, 2021.

[34] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. *arXiv preprint arXiv:2212.14532*, 2022.

[35] Tawfiq Salem, Menghua Zhai, Scott Workman, and Nathan Jacobs. A multimodal approach to mapping soundscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[36] Yi Wang, Conrad Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiaoxiang Zhu. Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 2022.

[37] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[38] Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[39] Rui Yang, Shuang Wang, Yingzhi Sun, Huan Zhang, Yu Liao, Yu Gu, Biao Hou, and Licheng Jiao. Multimodal fusion remote sensing image–audio retrieval. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:6220–6235, 2022.

[40] Ran Yue, Qi Meng, Da Yang, Yue Wu, Fangfang Liu, and Wei Yan. A visualized soundscape prediction model for design processes in urban parks. In *Building Simulation*, volume 16, pages 337–356. Springer, 2023.

[41] Tianhong Zhao, Xiucheng Liang, Wei Tu, Zhengdong Huang, and Filip Biljecki. Sensing urban soundscapes from street view imagery. *Computers, Environment and Urban Systems*, 99:101915, 2023.