Supplementary Material: Learning Tri-modal Embeddings for Zero-Shot Soundscape Mapping

BMVC 2023 Submission # 813

In this supplemental material, we present a demonstration of the zero-shot soundscape mapping capability offered by our proposed framework, GeoCLAP. Specifically, we showcase the soundscape maps created by querying our best performing model with diverse sound-related textual prompts. Furthermore, in a video demonstration accompanying this material, we highlight the satellite image to audio retrieval capability of GeoCLAP.

1 Zero-Shot Soundscape Mapping

Following the same methodology from Section 5.3 in the main paper, we constructed a soundscape map of England. We selected three prompts: This is a sound of car horn; This is a sound of chirping birds; This is a sound of animal farm. We downloaded sentinel2-cloudless images for England, each with dimension 256 × 256. Then, using cosine similarity scores between image and text embeddings, we created a dense soundscape map for the region. All visualizations were created using Q-GIS.

As observed in Figure 1, there is a strong correlation between sound categories and relevant land-cover classes. As expected, the soundscape map reveals that urban areas in England, such as the region around London, are highly associated with the sound category car horn indicated by the colour blue in Figure 1 (a). On the other hand, less populated areas with crops exhibit a notable association with the sound category animal farm. An intriguing observation is that around built-up areas in England, a combination of both car horn and chirping birds sound is observed, as indicated by purple-coloured regions in soundscape. This suggests that despite human activities in these regions, birds still inhabit them.

Soundscapes can be viewed as composite pseudo-colour maps representing a desired set of sound categories, as shown in Figure 1. However, if one is specifically interested in a single sound category, the GeoCLAP model can be queried with a textual prompt corresponding to that particular sound category, as demonstrated in Figure 2. Furthermore, visualizing soundscapes for smaller geographic regions, as showcased in Figure 3 and 4, can provide a better understanding of sound-related concepts learned by the model.

The results shown in Figure 3 indicate high similarity between the prompt: This is a sound of a manufacturing factory and a sub-region that likely contains structures resembling manufacturing factories. Similarly, in Figure 4, areas associated with water bodies exhibit a high similarity with the prompt: This is a sound of a flowing river. These findings demonstrate that the embedding space of GeoCLAP possesses an understanding of high-level sound-related concepts within geographic regions.

© 2023. The copyright of this document resides with its authors.
Figure 1: Comparison of (a) Soundscape map of England with (b) *ESRI’s 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*. Best viewed in colour.

Figure 2: Soundscape map of England created by querying GeoCLAP with a textual prompt: *This is a sound of church bells.*
Figure 3: Soundscape map of a small region (a) along with the reference overhead image (b). Soundscape created for the textual prompt: *This is a sound of manufacturing factory.* (green: more probable, white: less probable).

Figure 4: Soundscape map of a small region (a) along with the reference overhead image (b). Soundscape created for the textual prompt: *This is a sound of flowing river.* (green: more probable, white: less probable).