

AUDIO LATENT SPACE CARTOGRAPHY

Nicolas Jonason

KTH Royal Institute of Technology
njona@kth.se

Bob L. T. Sturm

KTH Royal Institute of Technology
bobs@kth.se

ABSTRACT

We explore the generation of visualizations of audio latent spaces using an audio-to-image generation pipeline. We believe this can help with the interpretability of audio latent spaces. We demonstrate a variety of results on the NSynth dataset.

1. INTRODUCTION

Many techniques exist for visualizing high dimensional latent spaces [1]. We present a new technique for providing visual support in the exploration of audio latent spaces. At a high level, this technique works by drawing audio samples from the latent space and using an audio-to-image generation pipeline to generate images which are then combined to form a map of the latent space. A web demo is available.¹

2. APPROACH

Figure 2 shows the audio latent space map creation process. The process involves the TimbreCLIP + Stable Diffusion audio-to-image pipeline described in [2]. At a high level, this pipeline interpolates the embeddings of multiple prepared text prompts based on how closely the input waveform relates to various keywords. The interpolated prompt embedding is then used to guide Stable Diffusion. [3] The keywords and template used to generate prompts are a crucial hyper-parameter which allow us to control how the audio latent space is depicted. Please refer to [2] for details about the audio-to-image generation pipeline.

3. DEMONSTRATION

Figure 1 shows a map created with our method. The latent space itself was constructed by applying the UMAP [4] algorithm to TimbreCLIP embeddings [2] of sounds from

¹<https://erl-j.github.io/audio-latent-space-cartography-demo>

© F. Author, S. Author, and T. Author. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).
Attribution: F. Author, S. Author, and T. Author, "Audio Latent Space Cartography", in *Extended Abstracts for the Late-Breaking Demo Session of the 23rd Int. Society for Music Information Retrieval Conf.*, Bengaluru, India, 2022.

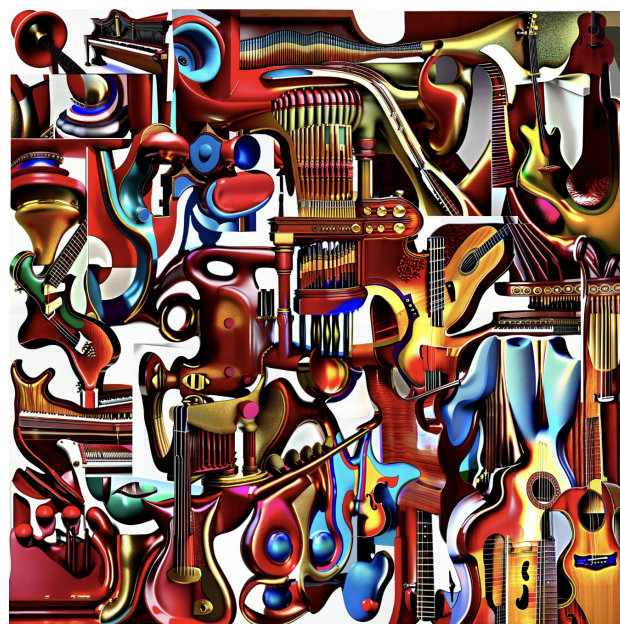


Figure 1. Map of a latent space containing 60 sounds from the NSynth dataset

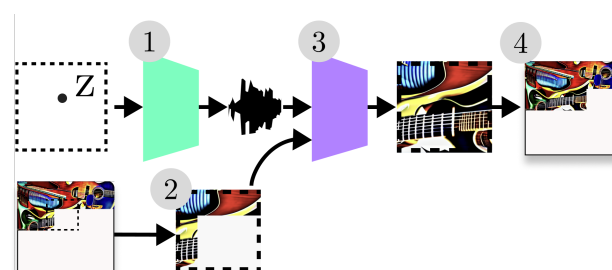


Figure 2. Diagram showing an iteration of the audio latent space map creation process. For each iteration, we draw latent coordinates in order to: 1) Generate audio; 2) Extract a patch of the current map to be inpainted. Then, we use an audio-to-image generation pipeline (TimbreCLIP + Stable Diffusion) to inpaint the patch (3). Finally, the map is updated with the inpainted patch (4)

the NSynth dataset [5]. We use 60 sounds from 10 different instrument families. All sounds were produced by playing the MIDI pitch C4. Figure 3 shows the same latent space map with the position and labels of the 60 samples superimposed onto the latent map.

Due to time constraints, we opted for the following simplification of the approach described in the previous sec-

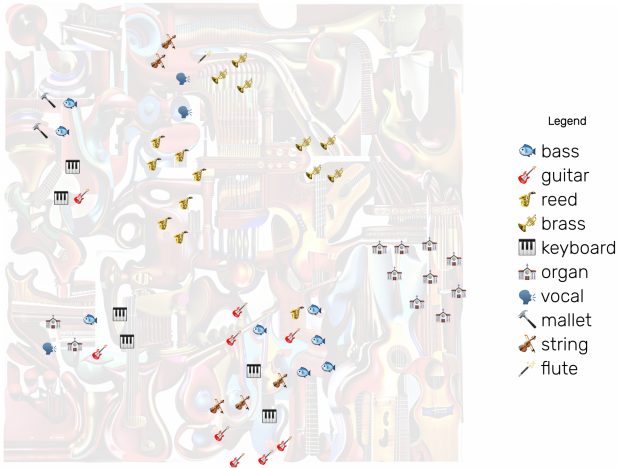


Figure 3. Latent map of 60 sounds from NSynth showing the coordinates of the 60 sounds in the latent space as well as their respective instrument families.

tion: instead of transforming every coordinate and then using that audio to drive the image generation, we instead generate a TimbreCLIP embedding using the inverse UMAP transformation and feed the TimbreCLIP embedding directly into the image synthesizer.

The prompt template used in the audio-to-image generation pipeline for Figures 1 and 3 is "A 3D rendered close-up of a <KEYWORD>, pinterest trending aesthetic". The keywords are a list of names of 21 musical instruments: "bass guitar", "acoustic guitar", "piano keyboard", "flute", "pipe organ", "violin", "cello", "double bass", "violin", "viola", "saxophone", "trumpet", "trombone", "tuba", "clarinet", "marimba", "kalimba", "xylophone", "bell", "electric guitar", "human voice". Maps using different prompt templates and keywords are showcased in Figure 4 as well as in the web demo.

4. ACKNOWLEDGMENTS

This paper is an outcome of MUSAiC, a project that has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant agreement No. 864189).



Figure 4. 4 maps of the same audio latent space in different styles. The map in the top left uses the style of the artist René Magritte, top right uses the style of the artist Zdzisław Beksiński. The map in the bottom left is in the style of Christmas ornaments, bottom right map depicts the audio latent space as mushrooms. Keywords and templates for these examples and more are available in the web demo.

5. REFERENCES

- [1] Y. Liu, E. Jun, Q. Li, and J. Heer, “Latent Space Cartography: Visual Analysis of Vector Space Embeddings,” *Computer Graphics Forum*, vol. 38, no. 3, pp. 67–78, Jun. 2019. [Online]. Available: <https://onlinelibrary.wiley.com/doi/10.1111/cgf.13672>
- [2] N. Jonason and B. L. T. Sturm, “TimbreCLIP: Connecting Timbre to Text and Images,” Nov. 2022, arXiv:2211.11225 [cs, eess]. [Online]. Available: <http://arxiv.org/abs/2211.11225>
- [3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” Apr. 2022, arXiv:2112.10752 [cs]. [Online]. Available: <http://arxiv.org/abs/2112.10752>
- [4] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” Sep. 2020, arXiv:1802.03426 [cs, stat]. [Online]. Available: <http://arxiv.org/abs/1802.03426>
- [5] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, “Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders,” in *Proceedings of the 34th International Conference on Machine Learning*. PMLR, Jul. 2017, pp. 1068–1077, iSSN: 2640-3498. [Online]. Available: <https://proceedings.mlr.press/v70/engel17a.html>